

SPEAKER RECOGNITION IN UNCONSTRAINED ENVIRONMENTS

ANDREAS NAUTSCH

Dissertation

Zur Erlangung des akademischen Grades
Doctor rerum naturalium
(Dr. rer. nat.)

genehmigte Dissertationsschrift in englischer Sprache
von Andreas Nautsch, M.Sc.
geboren in Greifswald

Erstreferent: Prof. Dr. rer. nat. Max Mühlhäuser
Korreferent: Prof. Dr.-Ing. Christoph Busch
Korreferent: Prof. Dr. Didier Meuwly

Tag der Einreichung: 28. Mai 2019
Tag der Prüfung: 10. Oktober 2019



Fachgebiet Telekooperation
Fachbereich Informatik
Technische Universität Darmstadt
Hochschulkennziffer D-17

Darmstadt 2019

Andreas Nautsch: *Speaker Recognition in Unconstrained Environments*,
Darmstadt, Technische Universität Darmstadt,
Jahr der Veröffentlichung der Dissertation auf TUPrints: 2019
Tag der mündlichen Prüfung: 10. Oktober 2019

Editorin: Elisabeth Neuhaus

Veröffentlicht unter CC BY-SA 4.0 International
<https://creativecommons.org/licenses>

ABSTRACT

Speaker recognition is applied in smart home devices, interactive voice response systems, call centers, online banking and payment solutions as well as in forensic scenarios. This dissertation is concerned with speaker recognition systems in unconstrained environments. Before this dissertation, research on *making better decisions in unconstrained environments* was insufficient. Aside from decision making, unconstrained environments imply two other subjects: security and privacy. Within the scope of this dissertation, these research subjects are regarded as both security against short-term replay attacks and privacy preservation within state-of-the-art biometric voice comparators in the light of a potential leak of biometric data. The aforementioned research subjects are united in this dissertation to sustain good decision making processes facing uncertainty from varying signal quality and to strengthen security as well as preserve privacy.

Conventionally, biometric comparators are trained to classify between mated and non-mated reference–probe pairs under idealistic conditions but are expected to operate well in the real world. However, the more the voice signal quality degrades, the more erroneous decisions are made. The severity of their impact depends on the requirements of a biometric application. In this dissertation, quality estimates are proposed and employed for the purpose of making better decisions on average in a formalized way (quantitative method), while the specifications of decision requirements of a biometric application remain unknown. By using the Bayesian decision framework, the specification of application-depending decision requirements is formalized, outlining operating points: the decision thresholds. The assessed quality conditions combine ambient and biometric noise, both of which occurring in commercial as well as in forensic application scenarios. Dual-use (civil and governmental) technology is investigated. As it seems unfeasible to train systems for every possible signal degradation, a low amount of quality conditions is used. After examining the impact of degrading signal quality on biometric feature extraction, the extraction is assumed ideal in order to conduct a fair benchmark. This dissertation proposes and investigates methods for propagating information about quality to decision making. By employing quality estimates, a biometric system’s output (comparison scores) is normalized in order to ensure that each score encodes the least-favorable decision trade-off in its value. Application development is segregated from requirement specification. Furthermore, class discrimination and score calibration performance is improved over all decision requirements for real world applications.

In contrast to the ISO/IEC 19795-1:2006 standard on biometric performance (error rates), this dissertation is based on biometric inference for probabilistic decision making (subject to prior probabilities and cost terms). This dissertation elaborates on the paradigm shift from requirements *by error rates* to requirements *by beliefs in priors and costs*. *Binary decision error trade-off* plots are proposed, interrelating error rates with prior and cost beliefs, i.e., formalized decision requirements. Verbal tags are introduced to summarize categories of *least-favorable decisions*: the plot's canvas follows from Bayesian decision theory. Empirical error rates are plotted, encoding categories of decision trade-offs by line styles. Performance is visualized in the latent decision subspace for evaluating empirical performance regarding changes in prior and cost based decision requirements.

Security against short-term audio replay attacks (a collage of sound units such as phonemes and syllables) is strengthened. The unit-selection attack is posed by the ASVspoof 2015 challenge (English speech data), representing the most difficult to detect voice presentation attack of this challenge. In this dissertation, unit-selection attacks are created for German speech data, where support vector machine and Gaussian mixture model classifiers are trained to detect collage edges in speech representations based on wavelet and Fourier analyses. Competitive results are reached compared to the challenged submissions.

Homomorphic encryption is proposed to preserve the privacy of biometric information in the case of database leakage. In this dissertation, log-likelihood ratio scores, representing biometric evidence objectively, are computed in the latent biometric subspace. Conventional comparators rely on the feature extraction to ideally represent biometric information, latent subspace comparators are trained to find ideal representations of the biometric information in voice reference and probe samples to be compared. Two protocols are proposed for the the two-covariance comparison model, a special case of probabilistic linear discriminant analysis. Log-likelihood ratio scores are computed in the encrypted domain based on encrypted representations of the biometric reference and probe. As a consequence, the biometric information conveyed in voice samples is, in contrast to many existing protection schemes, stored protected and without information loss. The first protocol preserves privacy of end-users, requiring one public/private key pair per biometric application. The latter protocol preserves privacy of end-users and comparator vendors with two key pairs. Comparators estimate the biometric evidence in the latent subspace, such that the subspace model requires data protection as well. In both protocols, log-likelihood ratio based decision making meets the requirements of the ISO/IEC 24745:2011 biometric information protection standard in terms of unlinkability, irreversibility, and renewability properties of the protected voice data.

ZUSAMMENFASSUNG

Die biometrische Sprechererkennung findet Anwendung in Smart-Home-Lösungen, interaktiven Sprachdialogsystemen, Call Centern, Online-Banking und mobilen Zahlungsverfahren sowie in der forensischen Fallarbeit. Die vorliegende Dissertation konzentriert sich auf die biometrische Sprechererkennung bei unkontrollierbaren Einflussfaktoren. Vor dieser Dissertation war die Forschung zum *Fällen besserer Entscheidungen bei unkontrollierbaren Einflussfaktoren* unzureichend. Abgesehen von Betrachtungen zur Entscheidungsfindung beinhalten unkontrollierbare Einflussfaktoren zwei weitere Themenkomplexe: Sicherheit und Datenschutz. Im Rahmen dieser Dissertation werden beide Gebiete bezüglich der Sicherheit gegen Kurzzeit-Replay-Angriffe und der Wahrung von Privatsphäre im Hinblick auf mögliche Leaks biometrischer Daten betrachtet. Die oben genannten Forschungsthemen werden vereint, um einerseits das Treffen guter Entscheidungen trotz variierender Unsicherheit (aufgrund variabler Signalqualität) zu verbessern und andererseits die Sicherheit biometrischer Sprecherkennungssysteme zu härten. Die Privatsphäre wird gleichzeitig geschützt.

Normalerweise werden biometrische Mustererkenner trainiert, um zwischen gepaarten und nicht gepaarten Teilen biometrischer Referenzen und Proben unter idealen Bedingungen zu klassifizieren, aber es wird auch erwartet, dass diese Erkenner in der realen Welt gut funktionieren. Je mehr sich jedoch die Qualität von Sprachsignalen verschlechtert, desto häufiger werden Fehlentscheidungen getroffen. Dabei hängt die Folgeschwere der Fehlentscheidungen von den Anforderungen an eine biometrische Anwendung ab. In dieser Arbeit werden Qualitätsschätzer vorgeschlagen und eingesetzt (quantitative Methode), um im Schnitt (innerhalb eines formalen Frameworks) bessere Entscheidungen zu treffen, während die Spezifikationen der Entscheidungsanforderungen einer biometrischen Anwendung beliebig, aber fest sind. Durch den Einsatz des Bayes'schen Entscheidungs-Frameworks wird die Spezifikation der anwendungsabhängigen Entscheidungsanforderungen formalisiert. Darauf basierend wird ein Schwellenwert abgeleitet, anhand dessen Entscheidungen automatisiert gefällt werden können. Die betrachteten Qualitätsbedingungen kombinieren Umgebungs- und biometrische Störsignale, die sowohl in kommerziellen als auch in forensischen Anwendungsszenarien auftreten können. Es wird die (zivile und staatliche) Dual-Use-Technologie untersucht. Mehrere Qualitätsbedingungen werden betrachtet, da es nicht möglich erscheint, für jede mögliche Signalverschlechterung ein Erkennungssystem zu trainieren. Die Auswirkungen aufgrund von Signalqualitätsverschlechterung auf die

biometrische Merkmalsextraktion werden untersucht. Nach der Untersuchung wird diese Extraktion als ideal angesehen, um anschließend faire Benchmarks durchzuführen. Diese Dissertation schlägt Methoden zur Anwendung von Informationen über die Qualität in der (biometrischen) Entscheidungsfindung vor und untersucht diese. Durch die Verwendung von Qualitätsschätzern werden die Resultate eines biometrischen Systems (Vergleichswerte) normiert, um sicherzustellen, dass jeder Wert den geringsten-günstigsten Entscheidungskompromiss in seinem Wert kodiert. Die Anwendungsentwicklung wird von der Anforderungsspezifikation getrennt. Dies ist aufgrund des Bayes'schen Entscheidungs-Frameworks möglich: Risikoanalysen und maschinelles Lernen sprechen die gleiche Sprache, wenn Bayes Wahrscheinlichkeiten diskutiert werden. Dadurch werden sowohl die Klassenunterscheidung als auch die Kalibrierungsleistung über alle Entscheidungsanforderungen für reale Anwendungen verbessert.

Im Gegensatz zur Norm ISO/IEC 19795-1:2006 über die biometrische Performanzauswertung (Fehlerraten aus Beobachtungen) basiert diese Arbeit auf biometrischer Inferenz für probabilistische Entscheidungsfindung (in Anbetracht verschiedener a-priori-Wahrscheinlichkeiten und Kosten für verschiedene Fehlertypen). Diese Dissertation trägt zum Paradigmenwechsel von Anforderungen *durch Fehlerraten* zu Anforderungen *durch Annahmen von Wahrscheinlichkeiten und Kosten* bei. *Binary Decision Error Trade-off* (BET) Plots werden vorgeschlagen, die die Fehlerraten mit den Annahmen von Wahrscheinlichkeiten und Kosten in Beziehung setzen (in Bezug zu formalisierten Entscheidungsanforderungen). Verbale Annotationen werden eingeführt, um Kategorien von *Entscheidungen minimalen Vorteils* zusammenzufassen: Das Koordinaten-Design des BET-Plots folgt aus der Bayes'schen Entscheidungstheorie, sodass Entscheidungskompromisse formalisierter Annahmen auf den Achsen abgetragen werden. Empirische Fehlerraten werden grafisch dargestellt. Dies hat den Mehrwert, dass Kategorien von Entscheidungskompromissen in der Performanz-Darstellung eines Erkennungssystems mit verschiedenen Linienstilen abgetragen werden. Die Performanz wird nicht im beobachteten Raum, sondern im latenten Unterraum der biometrischen Klassifikation visualisiert. Dadurch können Veränderungen in den Entscheidungen bezüglich angenommener Wahrscheinlichkeiten und Kosten in exaktem Bezug zur empirisch evaluierten Performanz eines Erkennungssystems bewertet werden.

Die Sicherheit gegen Angriffe wird somit erhöht; konkret gegen Angriffe, in denen Audios in kurzen Zeitsegmenten aufgenommen und wieder abgespielt werden (eine Collage von Sound-Einheiten wie Phonemen und Silben). Der sogenannte *Unit-Selection-Angriff* wird auf den Daten der ASVspoof 2015 Challenge (englische Sprachdaten) untersucht, die den am schwersten zu erkennenden Angriff dieser Challenge darstellen. In dieser Arbeit werden Unit-Selection-

Angriffe für deutsche Sprachdaten erstellt, bei denen Support Vector Machine und Gauß'sche Mischmodell-Klassifikatoren trainiert werden, um Collage-Kanten in Sprachdarstellungen aus Wavelet- und Fourier-Analysen zu erkennen. Im Vergleich zu den ASVspoof 2015 Teilnehmern werden vergleichbare Ergebnisse erzielt.

Zum Schutz der Privatsphäre biometrischer Informationen im Falle eines Datenbanklecks wird homomorphe Verschlüsselung vorgeschlagen. In dieser Dissertation werden sogenannte Log-Likelihood Ratio Scores berechnet, die die biometrische Beweislast im latenten biometrischen Unterraum objektiv darstellen. Herkömmliche Komparatoren verlassen sich darauf, dass die Merkmalsextraktion biometrische Informationen ideal darstellt (unter der Annahme von hoher Signalqualität). Für die biometrische Erkennung unter variierender Signalqualität hingegen sind genauere Methoden notwendig. Biometrische Mustererkenner werden trainiert, um Entscheidungen im latenten biometrischen Unterraum zu treffen, d.h. nur die biometrischen Informationen aus Sprachreferenzen und -proben werden miteinander verglichen (und keine anderen Faktoren haben darauf Einfluss). In dieser Dissertation werden zwei Protokolle für biometrische Komparatoren der sogenannten *Probabilistischen Linearen Diskriminanz-Analyse* (PLDA) vorgeschlagen, im konkreten Fall für das Zwei-Kovarianz Vergleichsmodell, einem PLDA Sonderfall. Die Vergleichswerte werden in Form von Log-Likelihood Ratios berechnet. Zum Datenschutz finden Berechnungen im verschlüsselten Raum, basierend auf verschlüsselten Darstellungen der biometrischen Referenz und Probe, statt. Dadurch werden die in den Sprachproben übermittelten biometrischen Informationen im Gegensatz zu vielen bestehenden Lösungen geschützt und ohne Informationsverlust gespeichert. Das erste Protokoll schützt die Privatsphäre der Endnutzer und erfordert ein Schlüsselpaar aus einem öffentlichen und einem privaten Schlüssel für jede biometrische Anwendung. Das letztgenannte Protokoll schützt die Privatsphäre von Endnutzern und Herstellern biometrischer Erkener unter der Verwendung von zwei Schlüsselpaaren. Da die verwendeten biometrischen Erkener die biometrische Beweislast in einem *trainierten* latenten Unterraum berechnen, stellt auch das Unterraummodell selbst sensitive Daten für einen Hersteller dar. Der Schutz dieser sensiblen Daten ist ein zweites Ziel, weshalb ein zweites Schlüsselpaar notwendig ist. Beide Protokolle, die Log-Likelihood-Ratio-basierte Entscheidungen fällen, erfüllen die Anforderungen des Standards zum Schutz biometrischer Informationen, ISO/IEC 24745:2011, besonders in Bezug auf Unverkettbarkeit, Irreversibilität und Erneuerbarkeit von abgespeicherten, geschützten biometrischen Daten.

PUBLICATIONS

JOURNALS

- [1] A. Nautsch, D. Meuwly, D. Ramos, J. Lindh, and C. Busch, "Making likelihood ratios digestible for cross-application performance assessment," *IEEE Signal Processing Letters (SPL)*, vol. 24, no. 10, pp. 1552–1556, Oct. 2017, [Online] <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8025342> [Code] <https://doi.org/10.24433/C0.154591c8-9d3f-47eb-b656-3aff245fd5c1>, accessed: 2017-10-05.
- [2] A. Nautsch, A. Jimenez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez-Barrero, D. Petrovska-Delcrétaz, G. Chollet, N. Evans, T. Schneider, J.-F. Bonastre, B. Raj, I. Trancoso, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech and Language, Special issue on Speaker and language characterization and recognition: voice modeling, conversion, synthesis and ethical aspects*, vol. 58, pp. 441–480, Nov. 2019, [Online] <https://doi.org/10.1016/j.csl.2019.06.001> [SurveyTalk] <https://www.youtube.com/watch?v=mywNMwZfbDo>, accessed 2019-10-20.
- [3] A. Nautsch, D. Ramos, and D. Meuwly, "Binary-decision error tradeoff (BET) plots," *Manuscript*, 2019.

CONFERENCES

- [1] A. Nautsch, C. Rathgeb, R. Saeidi, and C. Busch, "Entropy analysis of i-vector feature spaces in duration-sensitive speaker recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 4674–4678.
- [2] A. Nautsch, R. Saeidi, C. Rathgeb, and C. Busch, "Analysis of mutual duration and noise effects in speaker recognition: Benefits of condition-matched cohort selection in score normalization," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2015, pp. 3006–3010.
- [3] A. Nautsch, R. Bamberger, and C. Busch, "Decision robustness of voice activity segmentation in unconstrained mobile speaker recognition environments," in *Proc. GI/IEEE Intl. Conf. of the Biometrics Special Interest Group (BIOSIG)*, 2016, pp. 135–146.

- [4] A. Nautsch, H. Hao, T. Stafylakis, C. Rathgeb, and C. Busch, "Towards PLDA-RBM based speaker recognition in mobile environment: Designing stacked/deep PLDA-RBM systems," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 5055–5059.
- [5] A. Nautsch, R. Saeidi, C. Rathgeb, and C. Busch, "Robustness of quality-based score calibration of speaker recognition systems with respect to low-SNR and short-duration conditions," in *Proc. Odyssey 2016: The Speaker and Language Recognition Workshop*, 2016, pp. 358–365.
- [6] U. Scherhag, A. Nautsch, C. Rathgeb, and C. Busch, "Unit-selection attack detection based on unfiltered frequency-domain features," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2016, pp. 2209–2213.
- [7] A. Nautsch, S. T. Steen, and C. Busch, "Deep quality-informed score normalization for privacy-friendly speaker recognition in unconstrained environments," in *Proc. GI/IEEE Intl. Conf. of the Biometrics Special Interest Group (BIOSIG)*, 2017, pp. 243–250.
- [8] A. Nautsch, D. Meuwly, D. Ramos, J. Lindh, and C. Busch, "Making likelihood ratios digestible for cross-application performance assessment," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, SPL presentation, 2018.
- [9] A. Nautsch, S. Isadskiy, J. Kolberg, M. Gomez-Barrero, and C. Busch, "Homomorphic encryption for speaker recognition: Protection of biometric templates and vendor model parameters," in *Proc. Odyssey 2018: The Speaker and Language Recognition Workshop*, 2018, pp. 16–23.

BOOK CHAPTER

- [1] A. Nautsch and C. Busch, "Voice biometrics: How the technology is standardized," in *Voice Biometrics*, C. Garcia-Mateo and G. Chollet, Eds., Manuscript, IET, 2019.

OTHER PUBLICATIONS

- [1] A. Nautsch, A. Lozano Diez, and D. Ramos, "Full/posterior PLDA in speaker recognition: Technical literature review," Hochschule Darmstadt, Universidad Autónoma de Madrid, Tech. Rep., 2016.
- [2] A. Nautsch, D. Ramos, J. González-Rodríguez, C. Rathgeb, and C. Busch, "Making decisions with biometric systems: The usefulness of a Bayesian perspective," in *Proc. NIST Intl. Biometric Performance Testing Conf. (IBPC)*, 2016.

FURTHER CONTRIBUTIONS

JOURNAL

- [1] A. Treiber, A. Nautsch, J. Kolberg, T. Schneider, and C. Busch, “Privacy-preserving PLDA speaker verification using out-sourced secure computation,” *Speech Communication*, vol. 114, pp. 60–71, Nov. 2019, [Online] <https://doi.org/10.1016/j.specom.2019.09.004>, accessed 2019-10-16.

CONFERENCES

- [1] M. Paulini, C. Rathgeb, A. Nautsch, H. Reichau, H. Reininger, and C. Busch, “Multi-bit allocation: Preparing voice biometrics for template protection,” in *Proc. Odyssey 2016: The Speaker and Language Recognition Workshop*, 2016, pp. 291–296.
- [2] K. Lee, V. Hautamäki, T. Kinnunen, A. Larcher, C. Zhang, A. Nautsch, T. Stafylakis, *et al.*, “The I4U mega fusion and collaboration for NIST speaker recognition evaluation 2016,” in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2017, pp. 1328–1332.
- [3] U. Scherhag, A. Nautsch, C. Rathgeb, M. Gomez-Barrero, R. Veldhuis, *et al.*, “Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting,” in *Proc. GI/IEEE Intl. Conf. of the Biometrics Special Interest Group (BIOSIG)*, 2017, pp. 149–159.
- [4] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, “The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding,” in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, Manuscript, 2019.
- [5] A. Nautsch, J. Patino, A. Treiber, T. Stafylakis, P. Mizera, M. Todisco, T. Schneider, and N. Evans, “Privacy-preserving speaker recognition with cohort score normalisation,” in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, Manuscript, 2019.
- [6] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, “ASVspoof 2019: Future horizons in spoofed and fake audio detection,” in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, Manuscript, 2019.

OTHER PUBLICATIONS

- [1] K. A. Lee, H. Sun, A. Sizov, G. Wang, T. H. Nguyen, *et al.*, “The I4U submission to the 2016 NIST speaker recognition evaluation,” in *Proc. NIST SRE workshop*, 2016.
- [2] A. Nautsch, S. Isadskiy, C. Rathgeb, and C. Busch, “HDA NIST SRE’16 systems (CRISP & I4U),” in *Proc. NIST SRE workshop*, 2016.
- [3] J. Lindh, A. Nautsch, T. Leinonen, and J. Åkesson, “Comparison between perceptual and automatic systems on finnish phone speech data (FinEval1) - a pilot test using score simulations,” in *Proc. Annual Conf. of the Intl. Assoc. for Forensic Phonetics and Acoustics (IAFPA)*, 2017.
- [4] A. Nautsch, U. Scherhag, S. Isadskiy, C. Rathgeb, and C. Busch, “da/sec ASVspoof2017 submission,” in *Proc. ASVspoof evaluation*, 2017.

*Information is not knowledge.
Knowledge is not wisdom.
Wisdom is not truth.
Truth is not beauty.
Beauty is not love.
Love is not music.
Music is the best.*

— Frank Zappa, 1979.

ACKNOWLEDGMENTS

Foremost, I would like to thank my promoters, Prof. Christoph Busch, Prof. Max Mühlhäuser, and Prof. Didier Meuwly for providing me the opportunity to research on the topics of my interest. I owe special thanks to Christoph for his guidance, support, patience, council, and wisdom. He is responsible for my first steps in the research world back in 2011, and he put enough trust and encouragement in me to start my doctoral research, self-organize my internships, seek collaborators new to his community, anticipate project acquisitions and management, and engage in international standardization activities. I would also like to thank my previous employer, atip GmbH, in particular Klaus Kasper, Herbert Reininger, and Martin Wagner, who guided my first steps in speech processing and were confident enough to send me to my first NIST evaluation, and especially to Anne Schönwandt, for all her support and encouragement during my bachelor's and master's period (and for the Yoga teachings at the time). After my research projects in the da/sec group at Hochschule Darmstadt, I was fortunate that Prof. Nicholas Evans hired me as a researcher to his audio security and privacy group at EURECOM, which was possible because of an industrial research project with Omilia, arranged by Themis Stafylakis. The time in the Nice area not only gave me serenity to reflect my dissertation drafts, I was also directly involved with the evaluation of the 2019 ASVspoof challenge as part of its organizing consortium. I'm extremely glad about the opportunity to revise my dissertation with Elisabeth, a very valuable editor spotting the *Fehlerteufel* which sneaked in during drafting (thanks Jannis, for establishing the contact).

I owe thanks to Rahim Saeidi for taking special interest in this work, supervising, and guiding me at the beginning of my doctoral studies, and to Daniel Ramos for bringing my understanding of the Bayesian decision framework forward (*¡BIBA!*). I was fortunate to visit foreign institutions as a guest researcher, and would like to express my gratitude to the ATVS group at the Universidad Autónoma de Madrid, the

Voxalys AB, and the computer vision lab at the University of Nottingham. I would particularly like to express my appreciation to Daniel Ramos, Joaquin Gonzalez-Rodriguez, Jonas Lindh, and Themis Stafylakis as to Yorgos Tzimiropoulos for hosting me. During these stays, I had the opportunity to meet lots of fellow doctoral students and excellent researchers: Alicia, los Rubenos, Patryk, Esther, Marta, Abhi, Joel, Maria, Adrian, Aaron, Jing, Keerthy, Kike, Joseph, and Charles.

I would like to thank the I4U consortium of 62 researchers from 15 collaborating sites, tackling the 2016 NIST SRE in a joint effort, to Kong Aik Lee for calling the consortium and hosting online meetings, and to Anthony Larcher for his support. For involving me in a collaboration of 13 authors on the performance assessment of morphing attacks and for our discussions on the suitability of relative to absolute vulnerability reporting, I would like to thank Ulrich. After these two collaborations, I was able to engage in another large project. To lead and collaborate a joint interdisciplinary effort on data privacy in speech data was a great experience, for which I would like to express my gratitude to the 22 co-authors with backgrounds in speech and audio processing, paralinguistics, legal studies, cryptography as well as image and audio based biometrics (a manuscript which would be impossible to write without the contribution of each co-author). The verbose yet concise way legal experts draft papers (Catherine and Els discussing via tracked changes Word documents) is still impressive to me. I'm glad about the way Jascha, Amos, and Abelino interactively engaged discussions. Thanks to Bhiksha for "poking" me to pursue this collaboration and for involving Isabel. For helping my understanding of cryptography grow, I would like to thank Thomas, Amos, Michael, and Melek. At EURECOM, Nick introduced me to the ASVspoof organization committee—Junichi, Tomi, Kong Aik, Héctor, Sahid, Xin, Ville, and Massimiliano—who trusted me with carrying out the detailed analyses on the physical access task of all submitted countermeasure systems (147 systems of 50 participating labs); thank you all for letting me jump in at the evaluation stage and for entrusting me to co-organize the special session at the 2019 IEEE Automatic Speech Recognition and Understanding workshop as well as for proposing me among the guest editors of the ASVspoof special issue of the Computer Speech and Language journal.

For the valuable discussions on performance evaluation, I would like to thank Niko Brümmer, Didier Meuwly, Elham Tabassi, Jim Wayman, and Raymond Veldhuis. Also, thanks to Niko for playing Frisbee with the folks from Brno and me in the waters of the beach in Les Sables d'Olonne; a truly memorable moment. For taking on my initiative and organizing the *Voice, Lips and Face: A concept workshop*, thanks to Niki, Anil, Oscar, and Finnian. Getting to know the Oxford Wave Research team (including interns) was great, bringing together the clients of this forensic service provider and academia

pushing the bounds of technology was insightful for getting to know the needs of their clients (I liked the down-to-earth yet fun interaction with them); the workshop organization and social events were marvelous. To Christoph, for proposing me as treasurer to the biometrics special interest group (BIOSIG) of the Gesellschaft für Informatik e.V. (GI), in whose executive committee I have been delighted to serve together with Alexander, Arslan, Christoph, Detlef, Heiko, and Victor-Philipp since 2015 (nothing big but still, thank you). Furthermore, I would like to thank the exceptional individuals in the SC 37 committee with the International Organization for Standardization and the International Electrotechnical Commission, providing ample of opportunities to discuss biometric matters. (Also, for the road trips and adventures off the meeting schedules.)

I would like to express my gratitude for the opportunity to work with competent partners in both BioMobile projects, encouraging me to pursue my research. I had the opportunity to collaborate with excellent colleagues from atip GmbH, the German federal criminal police office, AUTHADA GmbH, DEUDAT GmbH, usd AG, the association of German banks, and the Fraunhofer IGD, and the following individuals: Hermine Reichau, Michael Jessen, Sven Kloppenburg, Sandra Kübler, Marcel Wetzel, Christian Frei, Ralf Almon, André Nash, and Olaf Henniger. Especially, I would like to thank Hendrik Terstiege for his active communication as our partner of the Hessian state ministry for higher education, research and the arts, funding our projects. For their time and effort, I would like to thank the volunteers contributing to our BioMobile database collection, which would not have been possible without the assistance of Sergey, Natasha, Marjan, Jannis, Julius, Ulrich, and Soulya. For their valuable feedback in conceptualizing depending requirements, I would like to thank Michael Jessen, Hermine Reichau, Rahim Saeidi, David van der Vloed, and Andreas Heinemann.

In the da/sec research group at Hochschule Darmstadt, I have had the pleasure to share time in work, breaks, achievements, stress, discussions, bad moods, and laughs with Martin, Anika, Ulrich, Marta, Jascha, Lorenz, Thomas, Jessica, Hareesh, Nicholas, Pawel, Sebastian, Lisa, Christian, Jannis, Daniel, Johannes, Chris, Reinhard, and Katharina. Especially, I would like to thank Ulrich for starting our lab tradition of Friday cakes. In this research group, we worked together with many student assistants and interns, Ahmed, Fabian, Henrik, Karina, and Ruben, to name just a few. Thanks to Christian for suggesting me as his substitute jury member to the 2018 CAST bachelor and master theses awards, an unexpected but relevant experience. As researchers in the era of large-scale databases and deep learning, we had quite increasing demands on our IT infrastructure, which the team around Sergio Vergata (Sandra, Steffen and Lars) satisfied marvelously with appropriate hardware, a highly scalable and distributed

file system, and virtualization in an IT security centered infrastructure. When I started at Hochschule Darmstadt in 2014, we joined the Linguistic Data Consortium, which continued throughout the BioMobile projects. I'm glad this initiative is jointly continued in funding by two faculties of Hochschule Darmstadt led by Prof. Stefan Rapp (whom I enjoyed the one or other coffee break and conversation on speech signal processing with). During my research time at da/sec, I had the pleasure to collaborate with, supervise, and learn from Hong, Reiner, Ulrich, Søren, and Sergey during their final master projects.

Having arrived at EURECOM, I am happy about the warm welcome I was fortunate to experience. The Moka coffees with Alison, Antonio, Chiara, Emanuele, Flavio, Héctor, Khawla, Lorenzo, Massimiliano, Pasquale, Pepe, Placido, Pramod, Thomas, and Valeria are neat breaks to escape from the daily routine. Thanks to my office mate, Sebastian. Coming from different groups of the security department, we share insights in discussions, the pondering about the code of others, and Thursdays as "Pasta day" with colleagues of the software and systems security group. My first weeks in France were enriched by the after work activities with many other researchers of the plenty groups at EURECOM, of which I would especially like to thank Marius, Robert, Giovanni, Xenofon, and Dimitrios. On the work side, I enjoyed being a part of the audio security and privacy group with their very interactive research style, especially during paper writing. Having Tomi visiting our group was a lucky coincident which fostered collaborations. I cannot express my gratitude to Omilia and Themos enough, in particular for taking the risks and costs in funding the project that kicked off my new position and for believing in me, allowing me to proceed with my research.

Finally, I am indebted to my friends and family, for all their support, joyful moments, encouragements, and understanding throughout this journey. For aiding me in the conquest of supplying Friday cakes to our lab; struggling together in pub quizzes; exchanging spiritual concepts from the Bahá'í Faith to Yoga; traveling through the Northwest of India, exploring Israel and the North of Italy (the latter destination was decided by coin toss in a rented car); and enjoying concerts, gaming, hiking, reunions as well as outdoor adventures.

To my dad, for our traditional festival camping at *Zappanale*, our family road trips through Scotland and the South of France, for helping me moving to the Nice area (driving forth and back from the Baltic to the Mediterranean Sea), and his endless embrace of kindness and balance: thank you.

CONTENTS

1	INTRODUCTION	1
1.1	Biometric Application Scenarios	2
1.1.1	Voice Verification in Daily Commercial Use . .	2
1.1.2	Governmental Use: Forensics	2
1.2	Scope and Motivation of this Dissertation	3
1.3	Motivated Research Questions of this Dissertation . . .	9
1.4	The Thesis Statement of this Dissertation	10
1.5	Outline of this Dissertation	10
1.6	Contributions of this Dissertation	12
2	FUNDAMENTALS	15
2.1	The Narrative: on Paradigms in Decision Making . . .	16
2.1.1	The Bayesian Decision Framework: Examples .	18
2.1.2	Likelihood Ratios: Well-Calibrated System Out-puts	24
2.2	Biometric Systems in ISO/IEC Standardization	25
2.2.1	Generalized System Design	26
2.2.2	Performance and Testing Reporting Framework	29
2.2.3	Presentation Attack Detection: Testing and Re- porting	31
2.3	Gap Analysis ISO/IEC 19795-1:2006	33
2.3.1	Proper Scoring Rules: in Brief	35
2.3.2	Frequentist and Bayesian: in a Nutshell	36
2.3.3	Thresholds According to Error Rates (Frequentist)	37
2.3.4	Thresholds before Evaluation (Bayesian)	41
2.4	Bayesian Decision Framework	44
2.4.1	Total Probability Theorem and Identity Inference	45
2.4.2	Decoupled Decision Layer	50
2.4.3	Decision Risk Performance	54
2.4.4	Error Rate Performance Visualization and Ideal Score Calibration	58
2.4.5	On Performance Visualizations in Forensic Evaluation	61
2.4.6	Information Performance: System Contribution to Decision Making	64
2.4.7	Goodness of LLR Scores	66

2.5	Automatic Speaker Recognition	68
2.5.1	Brief Overview of Speaker Recognition Technology	69
2.5.2	Acoustic Feature Extraction in Brief	72
2.5.3	Conventional Signal Processing and Comparison	75
2.5.4	Embeddings: Feature Extraction	82
2.5.5	Comparators for i-vector and x-vector Embeddings	88
2.5.6	Score Normalization and Calibration: Improving Decision Making	95
2.6	Summary and Conclusion	98
3	EXPERIMENTAL FRAMEWORK	103
3.1	Evaluation Methodology	103
3.1.1	Organization of Data	104
3.1.2	Performance Criteria	105
3.2	Datasets and Protocols	105
3.2.1	Protocol of the 2013 MOBIO SRE	106
3.2.2	I4U Protocol of the 2012 NIST SRE	107
3.2.3	2013–2014 NIST i-vector SRE Protocol	112
3.2.4	GSDC and ASVspoof 2015 Datasets	112
3.3	Summary	113
4	ON THE PERFORMANCE PARADIGM SHIFT	115
4.1	Contribution: Angular Bayes Operating Points	117
4.2	Verbal-scaled Detection Error Tradeoffs	119
4.2.1	Making Likelihood Ratios Digestible: Verbal Scales	120
4.2.2	Contribution: Verbal Scales of Least-Favorable Decisions	122
4.2.3	Contribution: Verbal Scales in DET Plots	123
4.2.4	Contribution: Communicating Threshold Requirements	125
4.3	Contribution: Binary-Decision Error Trade-off (BET) Plots	128
4.3.1	Score Calibration: Properties and Implications	131
4.3.2	Modeling Bayesian Decision Policies (not Scores)	132
4.3.3	Tractable Decision Model Parameters	133
4.3.4	Plot Distances Revealing Log-Odd Trade-Offs	134
4.3.5	Generalization to Priors as well as Priors and Costs	136
4.3.6	Implications for Interrelations	136
4.3.7	Examples: Experimental and Use-Case	137
4.4	Contribution: Normalized Empirical Cross-Entropy	142
4.5	Contribution: Taxonomy to Performance Visualizations	143
4.6	Summary	147

5	ON THE INFLUENCE OF DURATION AND NOISE	149
5.1	Biometric Distinctiveness of Voice Samples	150
5.1.1	Feature Space Information	151
5.1.2	Contribution: Analysis of Feature Space Information	154
5.2	Analysis on the Segmentation of Voice Samples	157
5.2.1	Overview: Voice Activity Detection Algorithms	158
5.2.2	VAD Metrics in Speech and Speaker Recognition	161
5.2.3	Contribution: Decision Robustness Performance	163
5.3	Summary	167
6	ENHANCING DECISION INFORMATION	169
6.1	Score Normalization: Quality Adaptive Thresholds . .	170
6.1.1	Unified Audio Characterization Motivated Quality Vectors	172
6.1.2	Analysis: i-vector Pool Mean Shift	174
6.1.3	Contribution: Unconstrained Cohort based Score Normalization	175
6.2	Score Calibration Using Quality Estimates	178
6.2.1	Quality Measure Functions (QMFs)	179
6.2.2	Unified Audio Characterizations (UACs)	180
6.2.3	Contribution: Function of Quality Estimates (FQEs)	180
6.2.4	Analysis: Duration Impact on Clean Speech . .	182
6.2.5	Analysis: SNR Impact on Full Duration	183
6.2.6	Analysis: Combined Duration and SNR Effects	184
6.2.7	Analysis: Pooled Duration and SNR Levels . . .	186
6.2.8	Analysis: Calibration Robustness to Unseen Conditions	187
6.3	Deep Score Normalization with Quality Estimates . . .	190
6.3.1	Conventional Deep Neural Networks	191
6.3.2	Contribution: Deep Quality Informed Score Normalization	192
6.3.3	Robustness Analysis: Contained SNR Levels and Noise Types	194
6.4	Summary	195

7	PRESENTATION ATTACK SECURITY, PRIVACY, AND DATA PROTECTION	199
7.1	Detection of Unit Selection Attacks	200
7.1.1	Voice PAD: the ASVspoof 2015 Challenge	201
7.1.2	Contribution: Countermeasure on Sound Unit Transitions	202
7.1.3	Analysis: Detection of Unit-Selection Presentation Attacks	205
7.2	Privacy and Data Protection: Homomorphic Encryption	209
7.2.1	Outline: Biometric Information Protection for Speaker Recognition	210
7.2.2	Overview: Biometric Information Protection . .	211
7.2.3	Homomorphic Cryptosystems	212
7.2.4	Revisiting PLDA and Δ Cov Comparators in Speaker Recognition	216
7.2.5	Contribution: Privacy Architectures	217
7.2.6	Proof-of-Concept Study	221
7.2.7	Complexity Analysis	223
7.3	Summary	224
8	CONCLUSION	227
8.1	Main Results and Contributions	228
8.2	Future Perspectives	232
A	GREEDY PLDA-RBM AND MOBILE ENVIRONMENTS	235
A.1	PLDA with Restricted Boltzmann Machines (RBMs) . .	236
A.1.1	Restricted Boltzmann Machines	236
A.1.2	PLDA-RBM Algorithm	236
A.2	Contribution: Deep PLDA-RBM Designs	237
A.2.1	Greedy Architectures	238
A.2.2	Experimental Validation	239
A.3	Summary	241
B	PAV-LLR CALIBRATION: STEP-BY-STEP EXAMPLE	243
	Glossary	247
	BIBLIOGRAPHY	257

LIST OF FIGURES

Figure 1.1	General process flow of biometric verification	1
Figure 1.2	Dependence among dissertation chapters . . .	11
Figure 2.1	Bayesian decision framework	22
Figure 2.2	General design of a biometric system	28
Figure 2.3	Examples of steppy DETs with synthetic score distributions	30
Figure 2.4	Overview on attack points in biometric systems with depicted scope of ISO/IEC 30107-1	31
Figure 2.5	General biometric system composition with presentation attack detection	32
Figure 2.6	Total probability theorem example	46
Figure 2.7	Partitioning of the total probability	48
Figure 2.8	Examples of APE plots for synthetic score distributions	55
Figure 2.9	Examples of NBER plots	57
Figure 2.10	Example regarding DETs	59
Figure 2.11	EER as maximum minDCF	59
Figure 2.12	Example of PAV	60
Figure 2.13	PAV-LLR example	61
Figure 2.14	Limit Tippett plot example	63
Figure 2.15	Zoo plot example	63
Figure 2.16	Examples of ECE plots	66
Figure 2.17	Overview of speaker recognition with processings, groups of comparators, and timeline . . .	69
Figure 2.18	Overview on score processing	71
Figure 2.19	HMM example.	78
Figure 2.20	Architecture of end-to-end DNN extracting x-vector embeddings	87
Figure 2.21	Identity subspace model	90
Figure 3.1	Baseline performance on I4U evaluation set affected by mutual signal degradation	111
Figure 4.1	Operating points as linear combination in y-inverted ROC space	118
Figure 4.2	Contribution: visualizing C_{llr} in the ROC space	120
Figure 4.3	Comparison of verbal LR scales	123
Figure 4.4	Contribution: verbal scale of least-favorable decisions	124
Figure 4.5	DCF slopes in DET plots	124
Figure 4.6	Contribution: verbal scales encoded DET plot	125
Figure 4.7	Contribution: representative operating points per verbal band	127

Figure 4.8	Comparison of ROC axes scales	129
Figure 4.9	Big picture: the BET plot	130
Figure 4.10	Comparison of <i>logit</i> and <i>probit</i> scales	133
Figure 4.11	BET plot with verbal scales	134
Figure 4.12	BET plot example: AFIS performance	138
Figure 4.13	BET plot example: AFIS BETs	139
Figure 4.14	Proposed NECE diagram	142
Figure 4.15	Taxonomy on performance visualizations . . .	144
Figure 5.1	Estimating the lower bound of relative entropy	153
Figure 5.2	Comparison of feature and score domain relative entropy	155
Figure 5.3	Speaker subspace accumulation by duration .	156
Figure 5.4	VAD decision example under changing environmental conditions	163
Figure 5.5	VAD decision robustness	164
Figure 5.6	Sensitivity of VAD performances to different SNR levels	165
Figure 6.1	Concept: re-biasing thresholds depending on the quality of samples	171
Figure 6.2	Condition confusion matrix on q-vectors . . .	173
Figure 6.3	Multi-variate Student's t-test: i-vector mean value similarity among conditions	175
Figure 6.4	C_{llr}^{\min} and EER comparison of conventional AS-norm to oracle cohorts and the proposed pre-selection in extreme conditions	176
Figure 6.5	Pre-selected cohort subjects and conditions by unique selection	177
Figure 6.6	Baseline performance of an uncalibrated system on CROWD conditions	181
Figure 6.7	Comparison of class \mathcal{A} and class \mathcal{B} score distributions before and after calibration	182
Figure 6.8	C_{llr}^{mc} comparison of duration conditions	183
Figure 6.9	C_{llr}^{mc} comparison of different SNR conditions .	183
Figure 6.10	C_{llr}^{\min} comparison of $Q_{q-vector}$ on SNR conditions	184
Figure 6.11	C_{llr}^{mc} comparison of combined low quality conditions	185
Figure 6.12	C_{llr}^{mc} comparison for conditions of either short duration or low-SNR	185
Figure 6.13	C_{llr}^{mc} comparison of good quality conditions . .	186
Figure 6.14	C_{llr}^{mc} comparison of combined low quality conditions	188
Figure 6.15	C_{llr}^{mc} comparison for conditions of either short duration or low-SNR	188
Figure 6.16	C_{llr}^{mc} comparison of good quality conditions . .	189
Figure 6.17	Deep score normalization with quality estimates	193

Figure 6.18	Relative C_{llr}^{\min} change on deep score normalization	195
Figure 7.1	Structure of presentation attacks	201
Figure 7.2	Example of a bona fide speech signal and transitions	203
Figure 7.3	Example of a unit-selection speech signal and transitions	203
Figure 7.4	Spectrogram of human speech signal and transitions	204
Figure 7.5	Spectrogram of a unit-selection speech signal and transitions	204
Figure 7.6	DET and BET plots for configurations with best EER on validation set	207
Figure 7.7	DET and BET plots for configurations with best EER on the ASVspoof set	208
Figure 7.8	Architecture of homomorphically encrypted cosine similarity comparison	216
Figure 7.9	Contribution: architecture of homomorphically encrypted PLDA/2Cov comparison solely protecting subject data	219
Figure 7.10	Contribution: architecture of protected biometric information and comparison model hyper-parameters	221
Figure 7.11	DET and BET comparison of the baseline PLDA/2Cov system and the proposed HE PLDA/2Cov schemes	222
Figure A.1	Factorization concepts	237
Figure A.2	Comparison of proposed deep PLDA-RBM designs	238
Figure A.3	Comparison of different numbers of hidden speaker and channel factors	239
Figure B.1	Examples of the PAV-LLR algorithm	244
Figure B.2	Example of Laplace's rule of succession impacting the visualization of ROCCHs	246

LIST OF TABLES

Table 2.1	Examples of Bayesian credible interval compared to Frequentist confidence interval	43
Table 3.1	Partitioning of the MOBIO database	106
Table 3.2	Overview on NIST SRE datasets	108
Table 3.3	Overview on the I4U calibration sets	109
Table 3.4	Label scheme for mutual duration and noise conditions	110
Table 3.5	Database partitioning on GSDC and ASVspoof 2015	113
Table 4.1	Verbal scales for likelihood ratios	121
Table 4.2	Centers of $C_{llr}^{ratio}(\eta)$ gravity on the scale of conclusion	126
Table 5.1	Relative entropy and performance comparison	154
Table 5.2	VAD algorithm performance comparison . . .	164
Table 6.1	QMF and FQE comparison of pooled conditions	187
Table 6.2	QMF and FQE robustness comparison of pooled conditions	187
Table 6.3	Benchmark of relative C_{llr}^{\min} changes to PLDA baseline	193
Table 7.1	Configuration for best observed EER	206
Table 7.2	Best observed configurations evaluated with validation and ASVspoof sets	207
Table 7.3	Complexity analysis for the Euclidean and cosine comparators	216
Table 7.4	Contribution: complexity analysis for the proposed PLDA/2Cov HE schemes	223
Table A.1	Performance of baseline systems on development set	239
Table A.2	C_{llr}^{\min} comparison of stacking concepts for hidden speaker unit extraction	240
Table A.3	C_{llr}^{\min} comparison of recovered i-vectors by the channel-stacked PLDA-RBM architecture . . .	240
Table A.4	HTER and C_{llr}^{\min} comparison of best single systems of the 2013 MOBIO SRE	241

ACRONYMS

2Cov	two covariance model
APCER	attack presentation classification error rate
APE	applied probability of error
ASR	automatic speech recognition
BDF	Bayesian decision framework
BET	binary decision error trade-off
BPCER	bona fide presentation classification error rate
DCF	decision cost function
DET	detection error trade-off
ECE	empirical cross-entropy
EER	equal error rate
EM	expectation-maximization algorithm
FAR	false accept rate
FMR	false match rate
FMR ₁₀₀	FNMR at a 1% FMR
FNMR	false non-match rate
FQE	function of quality estimate
FRR	false reject rate
FTA	failure-to-acquire rate
FTE	failure-to-enrol rate
GFAR	generalized false accept rate
GFRR	generalized false reject rate
GMM	Gaussian mixture model
HE	homomorphic encryption
i-vector	intermediate-sized vector

LLR	log-likelihood ratio
LR	likelihood ratio
MFCC	mel-frequency cepstral coefficient
minDCF	minimum DCF
minECE	minimum ECE
NBER	normalized Bayes error rate
NECE	normalized ECE
NIST	US National Institute of Standards and Technology
PAD	presentation attack detection
PAI	presentation attack instrument
PAIS	presentation attack instrument species
PAV	pool adjacent violators algorithm
PLDA	probabilistic linear discriminant analysis
QMF	quality measure function
q-vector	quality vector
ROC	receiver operating characteristic
ROCCH	ROC's convex hull
SNR	signal-to-noise ratio
SRE	speaker recognition evaluation
UAC	unified audio characterization
UBM	universal background model
VAD	voice activity detection

INTRODUCTION

Biometric speaker recognition is an emerging market, growing alongside the rising popularity of mobile end-user devices. In contrast to conventional authentication methods based on knowledge or tokens, biometric recognition aims at sustaining security while providing higher levels of convenience, e.g., passwords can be forgotten or cards can be lost, whereas biometric characteristics cannot. Based on behavioral and biological characteristics, biometric recognition systems extract and compare features in order to distinguish between individuals [1]. As speech is present in daily life, biometric voice authentication is utilized in many interactive application scenarios, e.g., contact centers and online banking, but also by governmental organizations to provide evidence in court or track suspects in investigations, among others. Capturing natural and free speech, it becomes relevant to address changes in environmental conditions, especially when it comes to noise degrading the biometric recognition performance. As mobile devices are increasing in availability, speaker recognition systems, in which voice samples are being captured, are increasingly facing unconstrained environments.

Biometric applications compare probe samples to previously enrolled biometric references with verification and identification recognition tasks. Identification systems either narrow down a short list of matching subjects or identify an individual as an element of a list of subjects, e.g., a visa black list in border control [2] or a member list of recent visitors with time-limited access [3]. Verifications are one-to-one comparisons, stating a binary decision problem of whether or not reference and probe samples stem from the same source, e.g., identity claims in banking [4] or forensic evidence reporting on the comparison of specimens and control samples [5]. Among others, biometric application domains are in payments, aviation, border control, and mobile devices with emerging domains in health care and privacy. The general process flow of biometric verification is depicted in Fig. 1.1: biometric features are extracted from segmented reference and probe samples and compared yielding a similarity score, such that an *accept* or *reject* decision is made by comparing a score to a threshold.

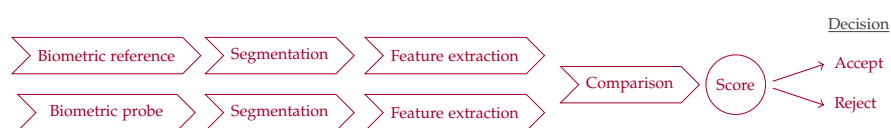


Figure 1.1: General process flow of biometric verification.

1.1 BIOMETRIC APPLICATION SCENARIOS

Speaker recognition is dual-use technology. Biometric voice comparison provides a convenient level of security in commercial application scenarios. It may furthermore provide evidence about identities in governmental scenarios.

1.1.1 *Voice Verification in Daily Commercial Use*

Commercial speaker recognition applications operate in stationary and mobile environments. Stationary environments are outdoor, e.g., facility access control gates [6, 7]; indoor, e.g., at home as well as with smart home applications like *Amazon Alexa* [8, 9] or authentication for e-learning platforms [10], office voice interactive help desk and password reset services [11], and at industrial workplaces. By contrast, mobile environments are shaped by smartphone based use cases, e.g., fraud prevention in contact centers [12], banking and payments applications [13] as well as smartphone voice authentication on Android [14, 15] and iOS devices [16]. Research in commercial applications concerns robust speaker modeling, the assessment of natural speech (text-independent recognition) and passphrases (text-dependent recognition), and the detection of [presentation attacks](#). Privacy-preserving technology is in high demand [17], speech services rely on data annotation which possibly includes sensitive information (e.g., names and bank accounts).¹

1.1.2 *Governmental Use: Forensics*

In forensic speaker recognition [5], case dependent specimen, e.g., voice probe samples, are compared with text-independent methods to controlled reference samples, which can stem from police interviews or other taped conversations. Usually, reference voice samples comprise natural, unaltered speech. However, probe voice samples might have been altered for disguise purposes, e.g., in blackmail calls or terrorist videos. Conventionally, forensic voice biometrics comprises manual, semi automated, and fully² automated methods, whereas linguistic and acoustic phonetic methodologies are the most widely

¹ Applications for speech transcription are freely available to the layman (e.g., see: <https://play.google.com/store/apps/details?id=com.google.audio.hearing.visualization.accessibility.scribe>, accessed 2019-04-25). The discussion on speech technology and privacy in daily use, however, needs to be addressed outside of this dissertation. First steps towards a common understanding of speech technology and legal communities are outlined in a collaborative work with Catherine Jasserand, Els Kindt, Massimiliano Todisco, Isabel Trancoso, and Nicholas Evans [18]. Existing technology safeguards are surveyed in a collaborative work with Bhiksha Raj, Thomas Schneider, and Christoph Busch (among others) [19].

² Fully automated methods are rarer. They possibly occur in surveillance scenarios.

spread form of evidence reporting. In linguistics, grammatical and stylistic patterns are examined for biometric characteristics in order to infer the biometric identity of a probe that is compared to a (closed-set) list of suspect identities³ (also to short-list), where phoneticians examine behavioral speech patterns and sound units, causing case dependent feature selection, comparison, and creation of a forensic report. Law enforcement speaker recognition software, such as *BATVOX*⁴, *Voice Biometrics GOV*⁵, *VoiceGrid*⁶, and *VOCALISE*⁷ is inspired from commercial applications as well as phonetic methodology, for which the selection and analysis of datasets is crucial to objective evidence reporting.

In practice [20], forensic science is based on four inferences: identification, individualization, association, and reconstruction. These inferences are structured in three levels: source level (origin of a trace), activity level (activity leading to the trace), and offense level (activity is constitutive of an offense). Currently, biometric systems computing scores are employed in three types of inferences at the source level: identification and identity verification, individualization, and association. The depending forensic biometric applications are: forensic identification, e.g., of disaster victims; forensic investigation, e.g., automated fingerprint identification systems; forensic intelligence, e.g., associating traces from different cases; and forensic evaluation, e.g., evaluation of biometric evidence in court by employing a biometric system of some sort. Forensic terminology differentiates between specimens as *probes* if captured in controlled conditions, e.g., existing in forensic identification; and as *trace* in uncontrolled (forensic) conditions. Research in forensic evaluation and investigation employing speaker recognition concerns limited data in training and testing. Applications may include large-scale suspect tracking.⁸ Additionally, language and environmental domain shifts cause decreasing biometric recognition performances that need to be compensated. Finally, fundamental principles on the quantification of the weight of evidence are under paradigm shift, i.e., a philosophical debate on *Bayesian* and *non-Bayesian* paradigms.

1.2 SCOPE AND MOTIVATION OF THIS DISSERTATION

The scope of this dissertation is to put biometric verification in applications according with the [Bayesian decision framework \(BDF\)](#). As exemplarily depicted, application fields are found in the commercial

³ Even if the operation is an open-set classification most of the time.

⁴ AGNITIO, Madrid, Spain, part of Nuance Communications, Burlington, MA, USA.

⁵ Phonexia, Brno, Czech Republic.

⁶ Speech Technology Center Ltd., St. Petersburg, Russia.

⁷ Oxford Wave Research Ltd., Oxford, UK.

⁸ See EU project: *Speaker Identification Integrated Project (SIIP) — Privacy Enhanced Speaker Identification at Global Reach*, [Online] <http://siip.eu>, accessed 2017-10-05.

as well as in the governmental sector. Therefore, the BDF is examined regarding the setup of an operating point (decision threshold based on beliefs), e.g., for communication between commercial entities and the communication between forensic experts and triers of fact. This dissertation distinguishes between commercial entities as biometric service providers, offering an authentication application to end-users; commercial entities as [biometric system operators](#), hosting and implementing a biometric service for a provider; and commercial entities as [biometrics \(sub-\) system vendors](#). The motivation of this dissertation is driven by the following observations from state-of-the-art speaker recognition, grouped by:

- different perspectives on *what performance is*;
- impact of changing environmental conditions;
- quality estimates from acoustic features;
- security, privacy, and data protection.

Different perspectives on *what performance is*. Performance evaluation constraints different figures of merit in speaker recognition [21, 22], forensics [23, 24], and biometrics standardization [25] research communities. In speaker recognition, application-depending operating points (thresholds) are parameterized by beliefs in the prior probability ratio of expected non-mated to mated comparison trials, and in the ratio of costs associated to falsely declaring a reference–probe pair "match" or "non-match". By weighting these beliefs by the empirical misclassification error rates of a system, the [decision cost function \(DCF\)](#) optimization criterion represents the expected cost of decision making within the BDF when employing a system. In the BDF, error rates are measured at an operating point, the Bayes threshold, resembling from prior and cost beliefs. By calibrating system score outputs to [log-likelihood ratio \(LLR\)](#) scores, DCFs are minimized (for the magnitude of belief ratios). Consequently, the empirical decision risk is reduced. Therefore, *discrimination* of recognition classes and *calibration* to produce scores that are LLRs *resemble performance*.

By contrast, concurrent biometric standardization solely refers to examining error rates, i.e., discrimination performance. The optimization of a recognition system to meet a prior and cost based decision policy is segregated (and unconcerned). Therefore, when employing a recognition system, discrimination is optimized without consideration of the impact of poor decision making. However, when solely developing and evaluating recognition systems for *one* operational scenario, both perspectives correspond to another (to an extent).

In forensics, evidence is reported objectively, and the provinces of the forensic practitioner and the court are separated.⁹ Automated

⁹ By segregating provinces of evidence reporting and jurisdiction, forensic practitioners and courts preserve their integrity within legal systems.

forensic systems report the weight of evidence (in this dissertation: LLRs), putting the magnitude of similarity and dissimilarity of the reference–probe pair into relation, i.e., prosecution and defense propositions. However, cost beliefs remain unparameterized in forensics, and the parameterization of the prior belief is in the province of court. Automated forensic systems are thus calibrated to provide information discriminative and well-calibrated enough to making good decisions on average, without knowing the prior beliefs set during an investigation. From speaker recognition and forensics, the figure of merit C_{llr} [26] emerged, application-independently summarizing discrimination and calibration as well as information performance. Among others, related work is contributed by the dissertations of Daniel Ramos [27], Niko Brümmer [28], and Rudolf Haraksim [29].

In order to optimize speaker recognition systems technology-independent of their use, a taxonomy on performance perspectives is required. Eventually, an interrelation of these perspectives provides guidance to the discourse of harmonization between the fields.

Impact of changing environmental conditions. Conventionally [21, 30, 31], speaker recognition systems are trained under idealistic conditions but are employed in the real world under different environmental influences, e.g., forensic in contrast to commercial scenarios, and operate under different decision policies. As voice reference samples are captured in controlled environments of rather ideal conditions, biometric voice probes are captured in real world situations.¹⁰ Thereby, the duration of voice samples is an indicator of the phonetic variety comprised, i.e., longer durations provide more sufficient biometric features as more speech is present (*sample completeness*).¹¹

On varying levels of interfering noise and different noise types, this dissertation concerns changing environmental constraints of two voice samples being compared, i.e., overlapping sounds stemming from heating and ventilation systems represent non-biometric (ambient) noise, and interfering voices of third speakers account for biometric noise, such that the biometric recognition task also needs to address third individuals recorded, e.g., from radios or TV sets running in the background or nearby persons chatting (in a recorded telephone call). Both noise types are present in commercial and forensic application scenarios.

¹⁰ By contrast, access border control is operating under supervision, thus high-quality of biometric probes is assured as well. This application scenario is controlled as lower-quality data is rejected and probe samples are recaptured at the gate.

¹¹ Short durations cannot allow for a vast variety of phonetic content within a voice sample. In text-dependent application scenarios, i.e., repetition of passphrases, the phonetic content remains rather fixed, such that *subjects* are recognized based on their characteristic utterance of those phrases. However, in text-independent application scenarios, i.e., natural speech, short duration samples can consist of vastly diverging phonetic contents. By contrast, the distribution of phonetic sound units converges for longer sample durations [32].

In this dissertation, the main focus is placed on duration and noise (from air conditioning and third speakers). Both tasks are repeatedly posed to the speaker recognition community by the [speaker recognition evaluation \(SRE\)](#) series of the [US National Institute of Standards and Technology \(NIST\)](#) [33–36]. Over the last decades, the NIST SREs have provided thousands of hours of speech data for the training of biometric voice comparators. However, this extensive availability of labeled data is not always given, e.g., small medium enterprises developing applications for mobile devices might not have the resources for mobile device speech data, and the characteristics of forensic speech data differ case-by-case. The *MOBIO SRE 2013* [37] poses a limited training dataset for *mobile environments*, serving as an exemplary study in this dissertation on readjusting model assumptions for the purpose of increasing discrimination performance.

Quality estimates from acoustic features. The calibration of scores for various environmental conditions is proposed in [38, 39]. Both rebias¹² score adaptively to an audio characteristic or a measure of quality, where the additive calibration term summarizes a quality comparison of reference and probe samples. Both studies use NIST SRE data. The first uses the *PRISM*¹³ subset [40] of the 2008 and 2010 NIST SREs, the latter uses the *I4U*¹⁴ subset [41] of the 2012 NIST SRE [35].

In [38], [unified audio characterization \(UAC\)](#) is proposed, estimating the posterior probability of seven conditions on (English) speech data, characterizing general audio properties, namely (1) solely telephone speech; (2) telephone and microphone speech; (3) interview and microphone speech; (4) vocal effort speech, summarizing normal, low, and high effort; (5) language, summarizing English, Chinese, Russian, Arabic, and Thai; (6) noisy microphone speech with different [signal-to-noise ratio \(SNR\)](#) levels, summarizing 0 dB, 15 dB, and 20 dB SNR levels; and (7) reverberant microphone speech data, summarizing 0.3, 0.5, and 0.7 reverberation time. As speaker recognition systems extract acoustic features first and then project them to the biometric subspace, UACs are estimated based on acoustic features, which are also applicable to non-biometric recognition tasks, such as speech, language, and emotion recognition. UACs reflect audio characteristics from the perspective of the recognition system.

In the dissertation of Miranti Indar Mandasari [39], quality is modeled for score calibration by [quality measure functions \(QMFs\)](#) of the

¹² Let's assume scores are well-calibrated in ideal conditions, then the quality degrades, and scores differ in expectation as well as true value (bias). By re-biasing scores adaptively to quality, well-calibrated scores might be achieved (research of this dissertation).

¹³ The subset comprises challenging speech data, *promoting robustness for speaker modeling (PRISM)*.

¹⁴ For the 2012 NIST SRE, the I4U consortium consisted of nine *university* and *industry* sites from *four* continents, hence the acronym.

duration and the SNR level. Thereby, duration serves as a proxy measure to the phonetic variability; the known SNR level is used instead of an SNR estimate. On the one hand, studies on the known SNR level provide a fair benchmark in this setup, however, operative systems need to estimate the SNR level as the noise-free speech signal remains unknown. Duration and SNR degraded speech data is derived by truncation of long duration and high SNR (clean) voice samples and synthetically added noise by using the *Filtering and Noise Adding Tool (FaNT)*.¹⁵ Thereby, speech data is segmented by voice activity into *speech* and *non-speech* sequences; segmentation labels of clean data are used for synthetic noise data. Quality conditions comprise five durations (5, 10, 20, 40 seconds, and more) and 25 SNR conditions (0, 5, 10, 15, and 20 dB), where air conditioning and biometric noise types are derived separately but summarized per SNR level. Symmetric patterns appear for reference and probe sample quality.

In order to normalize and calibrate scores of speaker recognition systems well when targeting unconstrained environments, different aspects need to be considered. There is an immense variety of quality aspects in speech data, e.g., reverberation, noise, language, emotions, and duration. The scope of this dissertation is limited on duration and noise, particularly on speech data derived from the I4U dataset. Rather than summarizing non-biometric and biometric noise types, i.e., ambient noise from air conditioning and noise from other speakers, this dissertation investigates on the impact of each noise type. References are assumed to be captured in high quality, as the reference sample capturing is usually supervised (in one form or another). As for the biometric recognition task, the biometric noise is expected to cause more severe impacts. By using measures such as SNR to model quality, measurement errors mislead the following signal processing. By estimating quality from the signal processing perspective, e.g., by UACs, the impact of changing signal quality is modeled more naturally from and for a system's signal processing.

Security, privacy, and data protection. Biometric systems operating in end-user authentication and forensics are in constant danger of being attacked. Weak points are found in different stages of the biometric signal processing [42–44], e.g., *presentation attacks* at the capture of a sample, intrusion of the database storing biometric references, and exploitation of the biometric comparator. These examples outline the scope of this dissertation on security, privacy, and data protection.

In the speaker recognition community, *presentation attack detection (PAD)* has been a regular object of research since 2013 [45] in special sessions of Interspeech and Odyssey conferences, elaborating on the *automatic speaker verification (ASV) spoofing and countermeasures challenge* (ASVspoof) editions in 2015, 2017, and 2019 [46–48], among

¹⁵ [Online] <http://dnt.kr.hsnr.de/download/fant.tar.gz>, accessed: 2018-12-30.

others. The focus of ASVspoof 2015 [49] is placed on speech synthesis, voice conversion (morphing), and short-term replay attacks. The origins of five attacks are known for the development of PAD modules but those of another five attacks remain unknown. As the known presentation attacks targeted frequency amplitudes of speech signals but not frequency phase shifts, four presentation attacks of the evaluation set sharing this approximation are well detected by the systems submitted to the challenge. However, the short-term replay attack, the *unit-selection attack*, preserved phase information, and is thus the difficult task of ASVspoof 2015. In ASVspoof 2017 [50], audio replay attacks are emphasized: speech is captured in different environments with multiple sensors and replayed in environments of the same and of different properties by varying playback devices and sensors. The conditions of development and evaluation data subsets differ in the way presentation attacks are performed. Replay attacks can, in principle, be implemented by any layman. ASVspoof 2019 [51] extends the aforementioned challenges. Presentation attacks are, *inter alia*, based on the outcomes of voice conversion challenges, another research field in speech communication. In contrast to the 2015 and 2017 editions, where error rates served as the primary figure of merit, the 2019 edition evaluates submissions based on the *tandem expected cost of decision errors* (t-DCF) [52] derived from speaker recognition and PAD performance. In this dissertation, the scope of security is limited to the detection of unit-selection attacks.

Preserving privacy in speaker and speech characterization is an interdisciplinary topic, bringing together technology experts, legal researchers, and cryptographers [19]. In speech communication, existing cryptographic approaches are based on, e.g., secure two-party computation, searchable symmetric encryption, functional encryption, hardware-assisted security, differential privacy, and *homomorphic encryption* (HE). The vast majority of cryptographic methods rely on integer modulo operations, whereas biometric voice comparisons are based on floating point operations, computing LLR scores. In contrast to conventional biometric comparators, such as Hamming, Euclidean, and cosine distances, where the amount of cryptographic operations scales linearly with the feature dimension, the computational complexity increases for estimating LLRs. State-of-the-art voice reference–probe comparators rely on matrix multiplication. Matrix inversions and determinant computations are also involved when propagating the uncertainty of the feature extraction. The propagation of uncertainty throughout comparisons [53, 54] addresses *unconstrained environments* by changing the outline of the comparison subsystem. The latter is assumed fixed in this dissertation to a standard state-of-the-art comparator.

In this dissertation, the computation of LLRs as scores is paramount, allowing no information losses. The scope on privacy and

data protection is limited to HE for state-of-the-art voice comparators, particularly the [two covariance model \(2Cov\)](#) computing LLR scores by a subspace model, a special case of the [probabilistic linear discriminant analysis \(PLDA\)](#) family. Privacy is preserved by the HE of aggregated biometric data. Data protection is provided to [biometric system vendors](#) of PLDA/2Cov comparators by the HE of the subspace model.

1.3 MOTIVATED RESEARCH QUESTIONS OF THIS DISSERTATION

The following research question is posed in this dissertation:

Can quality mismatches be estimated at pre-comparison stages and aid the Bayesian identity inference in discrimination and calibration performance?

In the context of motivation, further research questions arise from the main question:

To which extent can the Bayesian paradigm on performance be interrelated with conventional error rate trade-off diagrams, and how can established performance visualizations be classified in the [BDF](#)?

Can the impact of environmental conditions in terms of biometric distinctiveness—as voice references aggregate—and the robustness of voice sample segmentation decisions be quantified before the comparison subsystem?

As security systems are subject to attacks, can text-independent audio replay attacks which are based on unit-selections of previously captured and rearranged speech sequences be detected?

Can data privacy and data protection be preserved while sustaining performance? On privacy and data protection:

- Within the framework of the 2016 (EU) *General Data Privacy Regulation* [55], considering biometric data as *sensitive*, is privacy preservable for [biometric capture subjects](#) while sustaining performance?
- Can the data of comparison subsystem vendors be protected, meaning parameters of generative comparison models, that are trained on vast data amounts, while preserving privacy of capture subjects and sustaining performance?

Annex A addresses robust classification on limited training data:

Not all vendors of comparison subsystems are able to develop systems on datasets providing thousands of hours of speech. Can PLDA classifiers thus be de-noised on limited mobile device speech training data?

1.4 THE THESIS STATEMENT OF THIS DISSERTATION

The proposed methodology utilizes novel measures, examining feature space properties regarding environmental quality aspects, estimating the likelihood of a voice sample belonging to quality conditions. These quality estimates are employed in order to increase the biometric discrimination during score normalization and to calibrate system scores, facing unconstrained environmental conditions. By the proposed score calibration, LLRs result, i.e., perturbed scores are transformed into LLRs in a quality informed fashion. Effectively, scores represent the one-dimensional aggregated information of the biometric comparison, in which quality estimates are employed to aggregate the information about arising uncertainty of the feature extraction due to diverging environmental conditions. LLR scores have a coherent interpretation within the BDF, such that error-depending risks as well as occurrence probabilities can be adapted to changes in environmental quality conditions.

The thesis statement developed during this research period is the following:

The aggregated information for voice biometric identity inference can be enhanced using qualitative feature space information and re-biasing decision thresholds adaptively to biometric comparisons.

1.5 OUTLINE OF THIS DISSERTATION

This dissertation is conventionally structured in fundamentals depicting background theory and generally related work, the experimental framework introducing theoretical and practical methods as well as a proposed extension to the theoretical canon, and research studies with experimental validation. The dependency among chapters is illustrated in Fig. 1.2.

This dissertation is structured as follows:

- **Chapter 1** introduces the topic of automatic speaker recognition for commercial and forensic purposes. It constitutes the motivation of this dissertation and defines its scope, outline, and contributions.

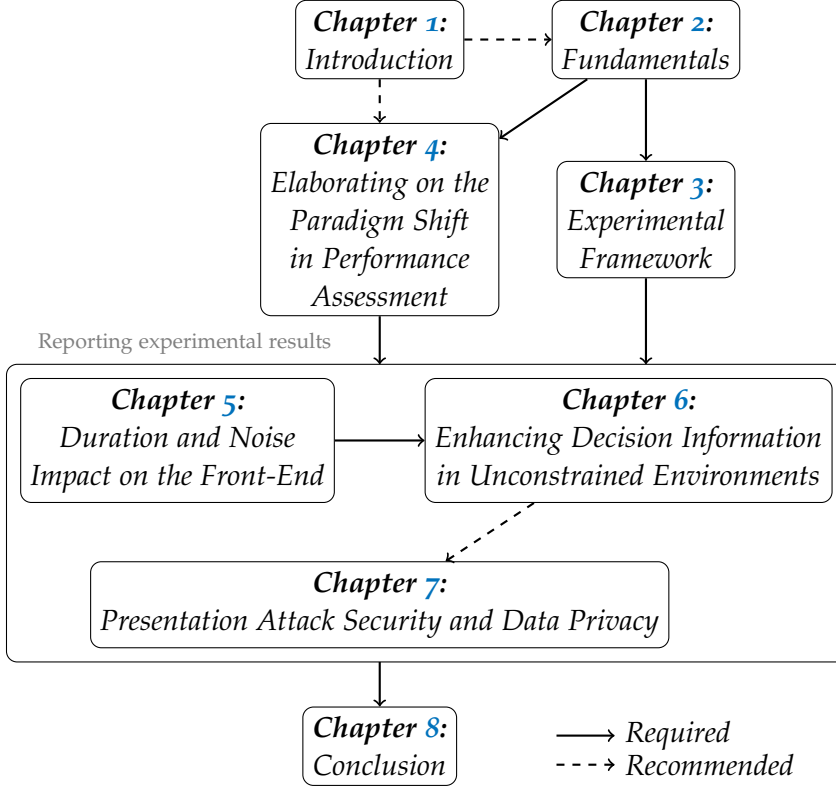


Figure 1.2: Dependence among dissertation chapters.

- **Chapter 2** summarizes related works in biometrics, standardization, BDF, and conventional signal processing as well as pattern recognition concepts in speaker recognition.
- **Chapter 3** describes the evaluation methodology, performance criteria, and evaluation databases.
- **Chapter 4** elaborates on the paradigm shift from Frequentist to Bayesian performance assessment, contributing visualizations.
- **Chapter 5** introduces a measure for the decision robustness of voice segmentation in the presence of noise and analyzes the biometric distinctiveness regarding variable speech durations.
- **Chapter 6** presents contributions regarding unconstrained environments, proposing quality adaptive score normalization and calibration methods, enhancing the biometric identity inference.
- **Chapter 7** addresses the security of speaker recognition systems in terms of detecting short-term replay attacks as well as privacy and data security aspects, proposing architectures for biometric cryptosystems, protecting privacy and data security demands alongside sustaining score calibration properties.
- **Chapter 8** concludes the dissertation, summarizing the main results and outlining perspectives for future research.

Chapters 2 to 6 are introduced by a preamble and concluded by a summary. For the purpose of seeking harmonization with the biometrics standardization community, ISO/IEC JTC1/SC37, this dissertation is written in compliance with the SC 37 terminology, but the language written is kept in American English (standards are written in British English). Equation readability is accommodated by indenting punctuation in the end of equations with a half space (to precaution misreadings, e.g., of commas after fractions with primes accenting denominator terms). Some methods developed in this dissertation are strongly based on popular approaches in speaker recognition [30, 31], Bayesian inference using probabilistic models [56–60], and BDF depending performance assessment methods [24, 26, 61, 62]. Methods developed in this dissertation are motivated by signal processing [63] and information theory [64]. This information is especially useful when dealing with Chapters 4, 5, and 6.

1.6 CONTRIBUTIONS OF THIS DISSERTATION

This dissertation is contributing the following:

To the Theoretical Framework

- definition of a visualization model consequently following the BDF interrelating error rates with depending decision log-odds, such that distances within the proposed **binary decision error trade-off (BET)** plot visually correspond to changes in BDF thresholds, representing belief changes in prior probabilities and decision costs [65];
- introduction of *verbal scales of least-favorable decisions* reflecting magnitudes of LLRs in error trade-off diagrams, making LLRs rather digestible from a Frequentist’s perspective and increasing the transparency of performance reporting in cross-application scenarios [66, 67];
- methodology on denoting thresholds (step-by-step) by refining priors and costs (exact value but also ranges) for commercial and jurisprudential communication, utilizing the BDF [66, 67];
- taxonomy on performance visualization for LLRs, distinguishing between i) the target audience (analysis or reporting), ii) criteria types (errors or information), and iii) criteria levels (discrimination or calibration), also proposing the **normalized ECE (NECE)** plot alongside. The taxonomy is self-contained;

To Novel Measures

- introduction of decision robustness of voice segmentation algorithms [68];
- outline of biometric voice distinctiveness: estimation of feature space entropy [69];
- definition of **quality vectors (q-vectors)** that linearly sample environmental condition parameters for quality informed score normalization [70], motivated by **UACs** [38] and **QMFs** [39];

To Novel Methods

- score normalization employing a quality-adaptive pre-selection of cohort data based on q-vectors for robust estimation of normalization parameters [70];
- score calibration for multi-condition environments, proposing **function of quality estimates (FQEs)** based on q-vectors [71];
- quality-informed score normalization, utilizing deep learning examining multi-condition environments [72];
- **PAD** of replay attacks as short-term audio collages in cross-language training and testing [73];
- HE for voice references of **biometric system end-users** and comparator models of **biometric system vendors** [74] (**Odyssey 2018 best paper award**);
- greedy learning of biometric information recovery during comparator estimation of feature subspaces [75], cf. annex A, when limited mobile device speech training data is solely available.

This chapter introduces the narrative of this dissertation on paradigms on decision making, followed by fundamentals of biometric systems, an outline of the [Bayesian decision framework \(BDF\)](#), and conventional approaches in automated speaker verification. Biometric systems are presented regarding general system design and definitions from ISO/IEC standardization within the JTC 1/SC 37 committee, such as harmonized vocabulary, performance assessment assuming zero-effort impostors, and a framework regarding the detection of presentation attacks. As the SC 37 perspective on performance solely considers a Frequentist's point of view while the speaker recognition community puts emphasis on fully Bayesian systems, i.e., following the BDF throughout, gaps between standardization and state-of-the-art speaker recognition communities are briefly depicted.

For the purpose of understanding the recognition metrics employed in this dissertation, relevant aspects of the BDF are explained, i.e., the total probability theorem, conditional probabilities, Bayesian inference, and the concept of decoupling comparison and decision layers, yielding a separation of *objective* score computation and *subjective* decision policies. Alongside, implications resulting from considering a Bayesian perspective are explained in order to provide a holistic overview on [log-likelihood ratio \(LLR\)](#) scores, highlighting benefits to machine learning for decision making. Notably, the decoupling of comparison and decision layers is encouraged by the forensic science community but is further applicable to commercial scenarios as well as to biometric standardization. By formalizing decision making, one can explicitly outline performance requirements (the values of score threshold) that systems need to meet (on low decision risks).

Finally, automated speaker recognition is introduced, particularly on the extraction of biometric features from acoustic data, the biometric comparison, and the adaptive normalization of comparison scores. The outlined baseline system depicts the 2012 to 2016 state-of-the-art feature extraction and comparison in speaker recognition, which is currently employed in products for commercial and forensic applications. Its *acoustic signal processing* extracts fixed-length [intermediate-sized vector \(i-vector\)](#) features from [mel-frequency cepstral coefficients \(MFCCs\)](#), which are a time-variable speech representation (since longer speech durations imply more MFCCs). By transforming acoustic features suitable to a variety of speech recognition problems (e.g., speech, language, and emotion recognition) into biometric features that are solely suitable to the biometric recognition task, com-

parisons are carried out in the *latent biometric subspace*. [Probabilistic linear discriminant analysis \(PLDA\)](#) comparators are state-of-the-art, carrying out comparisons in the latent biometric subspace, whereby the likelihoods of similarity and dissimilarity between reference and probe features are estimated as LLR scores. The family of PLDA comparators is suitable for $1 : 1$, $1 : n$ and $m : n$ recognition tasks. Systems based on the i-vector/PLDA framework are referred to as *fully Bayesian systems*; systems in consistency with the BDF. Eventually, score normalization and calibration techniques are employed in order to achieve higher biometric discrimination performance (low error rates) and to sustain well-calibrated scores (low decision risks). Such methods are relevant, as systems are trained on a distinct (ideal) domain but need to operate well in *unconstrained environments*.

2.1 THE NARRATIVE: ON PARADIGMS IN DECISION MAKING

Decision making¹⁶ with classification systems aims at performing the *best* possible action. Classification systems predict the class of observed input data, whereas prediction thresholds depend on the trade-off between wrong decisions.¹⁷ This is known to be best formalized by BDF [59], where an action must be taken towards a given *class*. In binary decision systems, two complementary propositions or classes exist.¹⁸ Commercial biometric applications might state:

- proposition *A*: *The biometric identity claim is true*; versus
- proposition *B*: *The biometric identity claim is false*;¹⁹

whereas forensic applications might state (e.g., for fingerprints [76]):

- proposition *A*: *same-source proposition: The fingerprint and the fingerprint originate from the same finger of the same donor*; versus
- proposition *B*: *different-source proposition: The fingerprint originates from a random finger of another donor of the relevant population, unrelated to the donor of the fingerprint*.²⁰

¹⁶ Parts of this section are based on a collaborative work with Daniel Ramos and Didier Meuwly [65].

¹⁷ In commercial applications, thresholds are defined by formal or informal requirements. In forensic applications, thresholds are denoted outside the scope of the forensic practitioner: decisions lie within the province of court. To assess the benefit of employing biometric systems within forensic evidence reports, decision making needs to be formalized in forensic evaluation, such that an information theoretic performance assessment can be conducted. The scope of this dissertation is to address commercial and forensic communities, leveraging formalized decision making.

¹⁸ The terms *proposition* and *class* will be used synonymously. In contrast to the term *hypothesis*, propositions are not implying randomness of the outcome.

¹⁹ In some other problems, both propositions can be non-exhaustive, yielding a multi-class problem. This discussion, however, exceeds the scope of this dissertation.

²⁰ Proposition *B* needs to be defined as *not A*, but both propositions might not reflect the entire reality. That is, the *relevant population* reduces the decision making problem to a tractable model for taking action on *A* over *B* or vice versa.

Decisions in commercial and governmental applications are made by following either a rule based (informal) or a risk/an information based (formal) paradigm. In both, erroneous decisions lead to Type I and Type II error rates. For the former paradigm, decision policies constrain absolute error rate trade-offs (e.g., 1% Type I error on Type II error rates below 5%). As such, a threshold (an application's operating point) is derived by examining the class score distributions of a system's evaluation output. In other words, operating points are derived *after* performance assessment based on informal specified requirements (error rate trade-offs). For the latter paradigm, decision policies are specified in a formal way, such that operating points are specified *before* performance assessment based on well-defined risk/information parameters: the probability of class \mathcal{A} , namely $\Pr(\mathcal{A})$ complementary to $\Pr(\mathcal{B}) = 1 - \Pr(\mathcal{A})$, and the cost of each erroneous decision $c_{\mathcal{A}}, c_{\mathcal{B}}$ (e.g., in a smart home application for lighting control, one could define $c_{\mathcal{B}} = \text{€}0.20$ for the power consumption cost caused by a neighbor activating light, whereas one could set $c_{\mathcal{A}} = \text{€}0.40$ for the inconvenience cost of not being recognized as oneself by the system). The emphasis of this dissertation is placed on the latter paradigm, the BDF.

In the BDF, the prior probability of class \mathcal{A} is updated by the observation of some findings E (e.g., biometric features and comparison scores). Ideally, the strength of the evidence for proposition \mathcal{A} versus \mathcal{B} is described by a so-called **likelihood ratio (LR)**, namely $\frac{\Pr(E|\mathcal{A})}{\Pr(E|\mathcal{B})}$, the likelihood of the evidence E under the model that assumes proposition \mathcal{A} (in ratio) to the likelihood of the evidence E under the model that assumes proposition \mathcal{B} . By updating the prior belief ($\Pr(\mathcal{A}), \Pr(\mathcal{B})$) with the LR, the posterior probabilities $\Pr(\mathcal{A}|E)$ and $\Pr(\mathcal{B}|E)$ of both classes are obtained.²¹ The posterior probability $\Pr(\mathcal{A}|E)$ is the likelihood of proposition \mathcal{A} to be *true* given the evidence E , and the posterior probability $\Pr(\mathcal{B}|E)$ is the likelihood of proposition \mathcal{B} to be *true* given the evidence E . Actions on \mathcal{A} are taken if the posterior ratio $\frac{\Pr(\mathcal{A}|E)}{\Pr(\mathcal{B}|E)}$ is equal to or exceeds the cost ratio that puts the cost of falsely accepting proposition \mathcal{B} in proportion to the cost of falsely rejecting proposition \mathcal{A} : $\frac{c_{\mathcal{B}}}{c_{\mathcal{A}}} \leq \frac{\Pr(\mathcal{A}|E)}{\Pr(\mathcal{B}|E)}$.

Considering the observation of some (biometric) features as the evidence E : information about the classes, the prior belief, is updated by the LR to the posterior belief. After obtaining posteriors, decisions are made following a well-known optimal decision rule, considering the costs $c_{\mathcal{A}}, c_{\mathcal{B}}$. In other words, LRs (computed by the system) are compared with the *Bayes threshold* $\frac{c_{\mathcal{B}}}{c_{\mathcal{A}}} \frac{\Pr(\mathcal{B})}{\Pr(\mathcal{A})}$; the joint ratio of prior probabilities and costs (external to the system).

²¹ As the propositions \mathcal{A}, \mathcal{B} are mutually exclusive (\mathcal{B} is *not* \mathcal{A}), the posterior probability $\Pr(\mathcal{A}|E)$ for the belief in proposition \mathcal{A} is derivable via the relationship $\frac{\Pr(\mathcal{A}|E)}{\Pr(\mathcal{B}|E)} = \frac{\Pr(\mathcal{A}|E)}{1 - \Pr(\mathcal{A}|E)}$. Notably, this relationship implies \mathcal{A}, \mathcal{B} being *exhaustive*, which holds for the *relevant population* within the decision model.

2.1.1.1 The Bayesian Decision Framework: Examples

Following the BDF paradigm, different perspectives on decision making are available, allowing to decouple system validation (development) and evaluation (operational) stages, as the gap in-between is bridged by sustaining well-calibrated scores (LRs). In the following examples, the effective ratios of costs and priors are summarized by the parameters c, π . Under the typical assumption of positive costs, the pair of (c, π) with the cost ratio $\frac{c}{1-c} = \frac{c_A}{c_B}$ and the prior ratio $\frac{\pi}{1-\pi} = \frac{\Pr(A)}{\Pr(B)}$ will define the whole range of possible applications of the system.²² As such, the cost parameter c is determinable by explicit costs (with monetary units). However, c remains without any unit, as one parameter value c summarizes various cost ratios, and π summarizes various prior ratios in a single value. The BDF decouples the subjective domain (formal requirements to decision making) from objective domain (a system's output as weight of evidence), such that one system is employable to various applications (c, π) and in different application fields. Depending on the (c, π) parameterization, different Type I and Type II error rates result. Type I errors are the class B reference–probe comparisons whose corresponding LR exceeds the (c, π) decision requirement, leading to a false action (erroneous acceptance of proposition B). Type II errors are the class A reference–probe comparisons whose corresponding LR fails to meet the (c, π) decision requirement, leading to a false action (erroneous rejection of proposition A). Ideally, LRs encode the overall proportion of erroneous decisions made at their depending Bayes threshold.

Example: Access Border Control (ABC) Gates

For face biometrics in ABC systems, FRONTEX requires a Type I error rate of 0.1%; the Type II error rate shall not to exceed 5% [77]. In [65], a collaboration contributing to the dissertation at hand, this is demonstrated so as to relate to explicit cost and prior values when following the BDF: assuming no prior intel ($\pi = 50\%$, known as *maximum prior entropy*), a cost value of $c \geq \frac{19}{19+999} = \frac{1}{1018} \approx \frac{1}{54}$ is required in order to meet the stated requirements—the 999 reflects the 0.1% Type I error rate requirement and the 19 the 5% Type II error rate requirement. Given a rather friendly environment (e.g., $\pi = 95\% = \frac{19}{20}$ of the biometric identity claims are true), a more security-demanding cost value is required instead: $c \geq \frac{19}{19+999 \times 19} = \frac{1}{1000}$. The error rate constraint remained fixed, but one aspect of the mutual belief in priors and costs changed, causing the 999 term to be multiplied with a 19 (in ratios, the denominator value 20 cancels out): in the light of a much more gen-

²² Linguistically, applications are defined in terms of ratios of proposition A to proposition B , especially as proposition B emerges from the terminology *not A*. In other words, expressions of the form $A : B$ are consistent within the BDF to expressions of the form $B : A$. Nonetheless, translating one to the other is trivial.

uine environment, the convenience centered yet secure cost constraint ($c \rightarrow 1$, here $\approx \frac{1}{54}$) becomes much more security centered one ($c \rightarrow 0$, here: $\frac{1}{1000}$). As the prior π in ABC gates can be fixed to some value, the criterion to which one biometric system is to favor another corresponds to the optimization of a security decision risk (lower c values are to meet). In this example, the mutual optimization of multiple (c, π) parameterizations remains out-of-scope. From the scope of a single **biometric system operator** with a single application, the BDF provides the same insights into which of two competing systems is preferable. Finding the best system under a fixed error rate constraint optimizes convenience levels at a fixed security level.

The rule based decision framework, however, solely provides an indicator of the *impact* of erroneous decisions. The benefit of the BDF unfolds in the assessment of cross-application performance. A particular application represents a possible scenario in which a system using LRs will be operating. An application's operating point is determined by its decision policy that is formalized by a particular pair (c, π) . Thus, a particular LR threshold value corresponds to one application scenario and many LR thresholds resemble many applications in a formalized manner. These values change across different application domains, for all of which LR scores satisfy the optimal decision rule, i.e., the Bayes risk: $\Pr(\mathcal{B} | E) c_{\mathcal{B}} \leq \Pr(\mathcal{A} | E) c_{\mathcal{A}}$.

Example: Research and Development (of Payment Services)

Biometric system vendors distributing the same biometric system to multiple biometric system operators need to meet different application domain requirements. For instance, sensitive bank applications with few impostors and many clients will require low c (penalizing false acceptances; $c_{\mathcal{A}}$ values become more negligible the larger $c_{\mathcal{B}}$ values become, effectively causing higher LR thresholds) and high π (lower probability of impostors). In contrast, a non-critical access control scenario with many impostors might require higher c (non-critical) and lower π (many impostors). Likewise, payment services concerning a broad range of transaction volumes will face different threat vectors, associating multiple (c, π) operating points depending on the decision risk. Thereby, varying security countermeasures (e.g., contactless payment, PIN, and signature verification) relate to specific (c, π) requirements and thus to the following thresholds: low LRs on contactless payment and high LRs on PIN or signature verification. In other words, depending on the LR value contributed by a biometric system, subsequent payment verification mechanisms can be chosen in the transaction authentication process. As application domains remain unknown during research and system development, vendors benefit from creating systems outputting LRs. The BDF allows to validate single binary

decision systems over all operating points (c, π) in a formal way: the parameterization of (c, π) outlines the Bayes threshold at which the empirical decision risk is measured, i.e., a (c, π) -weighted sum of Type I and Type II error rates. The Bayes risk defines performance requirements, i.e., an optimal decision risk. Targeting applications supporting multiple (c, π) parameterizations, e.g., varying cost levels depending on payment transaction volumes, a system's output needs to yield low Bayes risks independent of the application. These system output scores are known as LRs. Consequently, research on binary classification systems following the BDF can reach out to a broader audience as performance reporting is fully transparent within the formal decision making based on well-calibrated scores. Notably, scores of any (binary) classifier are transformable into (ideal) LRs [78] by *score calibration* [28, 79].

If scores are well-calibrated (LRs), they can be optimally used for all (c, π) parameterizations [26, 80]. However, if the scores are badly calibrated and the application (c, π) changes, new thresholds cannot be set without making new development experiments with new databases in order to empirically set optimal thresholds. Unfortunately, this cannot be done easily in many scenarios and in some it is simply impossible. For instance, in setups of varying security conditions, recalculating thresholds with a database each time a new security condition is established seems not feasible.

Example: Evaluation of Forensic Biometrics

Forensic experts and the courts address propositions on different levels: source and activity level for the experts and offense level for the court. Here, evidence reporting is decoupled from decision making. The goal of the forensic expert is to report the evidence for ideal decision making. The evidence reporting is thereby carried out on a different level than the decision making in court. The duty of the court is to provide the prior probability ratio and to make decisions based on the posterior probability ratio [20, 81]. The province of the forensic expert is restricted to the weight of evidence (the LR). In order to evaluate decision making, by employing one biometric system or another, the forensic expert can solely report on the information gain by employing a biometric system across all prior configurations. A solution to this problem consists in forcing systems to generate well-calibrated scores: LRs. LR scores can either result from the feature domain (fully Bayesian systems compute LRs by models assessing the biometric evidence) or from the score domain (calibration functions are trained that transform any score into their depending LR value). Systems that sustain LR scores meet the optimal Bayes risk at any Bayes threshold. By employing fully Bayesian systems, changing the threshold does not require the use of a development database but simply the

definition of priors π and costs c . As for commercial applications, this simplifies the change of applications (thresholds) enormously and makes technology application-independent (it adequately sustains decision making across all thresholds possible) [61, 80]. For forensic evaluation, (a) computing LR_s and (b) formally denoting thresholds are distinguished, accounting for the decoupled levels of the different provinces of the forensic expert (source and activation level) and of the court (offense level). In contrast to commercial scenarios, where priors are usually high (convenience), forensic scenarios consider rather low priors (e.g., $\pi = 1\%$ for one culprit among 100 suspects with prior odds 1 : 99) as one culprit is within a group of suspects. In commercials, this corresponds to security scenarios. Again, at the source and activation level, there is solely the estimation of the LR as the weight of evidence, whereas the prior parameterization and decision making is solely at the offense level. These decoupled provinces correspond in commercials to the biometric system computing LR_s (for source and activation levels) and the operative decision policy (for the offense level).

Notably, the employment of the BDF leverages the paradigm shift towards the Bayesian perspective: the LR score computation (province of the *expert witness*) is explicitly decoupled from decision making (province of the court). Ideally, the court would parameterize priors and costs, but this exceeds the scope of this dissertation. The automation of the jurisprudential process is explicitly *not targeted*: by providing well-calibrated scores (LR_s), the weight of evidence reported by the *expert witness* is objective and suitable for all operating points. In this sense, (c, π) values are solely parameterized theoretically, whereas commercial applications specify particular values, such that explicit thresholds are derived. In forensic evaluation, however, discussion about priors and costs is not applicable for the only focus is the strength of evidence. For the purpose of evaluating automated forensic biometric systems, LR_s are reported symmetrically, favoring *prosecution* and *defense* propositions, i.e., seeking an unbiased evaluation. One might link the example of forensic biometrics to the research and development phase of (binary decision) systems.

In this dissertation, the BDF is addressed under varying environmental conditions for speaker recognition systems. A system is trained on voice data stemming from ideal environmental conditions, a non-noisy environment with more than 40 s of captured speech. Yet, its operational environment varies as *subjects* are interacting with voice biometric systems in different settings, e.g., in offices with (heavy) ventilation/air conditioning (*ambient noise*) or at home with (voices from) television or the radio running in the background (*biometric noise*). Even in an unconstrained environment, ideal decision making is anticipated by end-users having expectations on the per-

formance of the system and by service providers sustaining low decision risks. The BDF formalizes optimal decision making (cf. Fig. 2.1). Thus, it is fundamental to address *speaker recognition in unconstrained environments*.

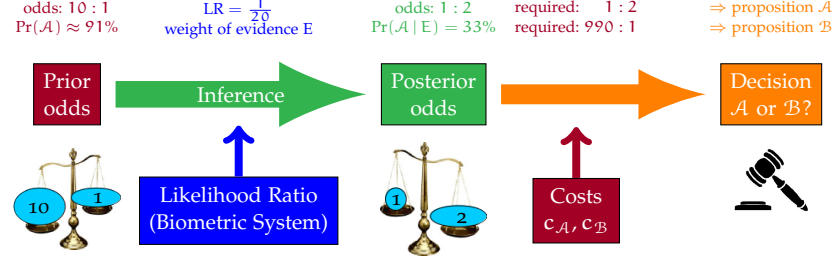


Figure 2.1: Bayesian decision framework, cf. [82], red: subjective operating point, blue: objective likelihood ratio reporting, green: chain of identity inference, orange: decision layer. Notably, in forensics, the judge adjudicates on the guilt, not on the identity of a subject. Similarly, in commercials, decisions are not made *about* an identity—although identity inference is used—; the granting of access is decided on.

Example: Smart Home, low Signal-Quality or false Claim?

The following example considers an application in an environment with few subversive interaction attempts, e.g., speaker recognition in smart home applications with high prior odds 10 : 1 ($\pi = \frac{10}{11}$); *true* biometric identity claims are ten times more likely to occur on average than *false* identity claims. Compared to a non-mated probe, a biometric system output in terms of evidence should result in low LR, i.e., supporting to favor proposition B over proposition A. In this example, a system compares a biometric reference with a biometric probe and weights the biometric evidence with a LR of $\frac{1}{20}$: $\Pr(E|B)$ is twenty times higher than $\Pr(E|A)$. A biometric system computes both probabilities and reports their ratio, such that a LR of $\frac{1}{20}$ resembles intermediate output ratios, such as $\frac{1\%}{20\%}$, $\frac{2\%}{40\%}$, $\frac{4.75\%}{95\%}$. Decisions are made considering the cost value of erroneous decisions c , which depends on the application policy. Smart home applications for lighting control (operated by speech) are non-critical, e.g., with a decision cost trade-off between 20 cents and 40 cents in power consumption versus inconvenience ($c = \frac{2}{3}$). A posterior ratio of at least $\frac{1-c}{c} = \frac{1}{2}$ is required (posterior odds 1 : 2), resembling a posterior belief in proposition A given the evidence of $\Pr(A|E) \geq \frac{\frac{1}{2}}{1+\frac{1}{2}} = \frac{1}{3} \approx 33.3\%$. While access to home security system administration is rather critical, e.g., with a decision cost trade-off in liabilities versus inconvenience of 99 000 € to 100 € ($c = \frac{1}{991}$), a posterior ratio of at least $\frac{990}{1}$ is required (odds 990 : 1 at $\Pr(A|E) \geq \frac{990}{991} \approx 99.9\%$).

The identity inference process following the BDF is illustrated in Fig. 2.1: the belief in the prior ratio of $\pi = \frac{10}{11}$ (odds 10 : 1 at $\Pr(\mathcal{A}) = \frac{10}{11} \approx 91.0\%$) is updated by an LR of $\frac{1}{20}$ to the belief in the posterior ratio of $\frac{\pi}{1-\pi} \frac{1}{20} = \frac{10}{1} \frac{1}{20} = \frac{1}{2}$ (odds 1 : 2 at $\Pr(\mathcal{A} | \mathcal{E}) = \frac{1}{3} \approx 33.3\%$). Thus, the convenient smart home application decides on proposition \mathcal{A} , whereas the security application decides on proposition \mathcal{B} .

However, the amount of captured voice signals (audio duration) varies among interaction instances, causing the biometric voice comparison to operate on mismatching sample completeness conditions. Also, facing increasing noise levels, e.g., originating from ventilation systems and television sets running in the background during a biometric capture process (ambient and biometric noise), the biometric system operates on mismatching noise conditions. Either way, the estimation precision of LR declines, effectively resulting in the miscalibration of the biometric system output: a score of $\frac{1}{20}$ will lead to misconducted decision making if not accounting for changes in, e.g., sample completeness and signal quality. Let's assume the biometric probe is mated, high well-calibrated scores are expected, such as an LR of $\frac{99}{1}$, but an LR of $\frac{1}{20}$ is observed. For a prior $\pi = \frac{10}{11}$, a posterior ratio of $\frac{\pi}{1-\pi} \frac{99}{1} = \frac{990}{1}$ resembles (at the posterior probability $\Pr(\mathcal{A} | \mathcal{E}) = \frac{990}{991} \approx 99.9\%$), meeting the requirements of either exemplary smart home application in order to decide for proposition \mathcal{A} . For the purpose of making good decisions on average, formal decision making requirements (c, π) need to be refined for a speaker recognition system in unconstrained environments, such that low quality is capable of justifying to decide for a true identity claim and against a false claim.

In this dissertation, LR values are re-biased by an offset, which depends on the sample quality. By employing a fully Bayesian system as baseline, well-calibrated scores (LRs) are sustained from the feature domain by employing appropriate signal processing and modeling. Facing unconstrained environments, scores (corresponding to ideal LRs on idealistic controlled conditions) become uncalibrated and are unsuitable to ideal decision making across all thresholds if the biometric voice sample quality degrades. This dissertation investigates on sustaining biometric discrimination (low error rates) and calibration (low Bayes risk across all thresholds) in unconstrained environments. Therefore, score normalization and calibration schemes are examined (based on estimates of biometric quality), producing discriminative scores and well-calibrated system outputs: LR scores.

2.1.2 Likelihood Ratios: Well-Calibrated System Outputs

Classification systems are well-calibrated if the yielded LR_s are representative of the feature distribution [26, 80, 83]. Thus, calibration is a property of scores: well-calibrated scores are LR_s. Following the Bayesian rule, decisions are made based on LR_s: optimal decisions are performed if a classifier yields scores comparable to the Bayes threshold [26, 80]. However, if the LR_s of the classifier are not well-calibrated, then the Bayes threshold will not be optimal either. Uncalibrated LR_s will be referred to as simply *scores* and well-calibrated scores as $\log(\text{LR})$ of **LLR_s**.²³ The calibration property is important for applications requiring a transparent relation between the strength of evidence and the decision risk as systems are used in any kind of application ranging from high-security (e.g., access to bank accounts) to high-convenience (e.g., access to personalized publicity). For the former, systems must be restrictive with impostors. For the latter, systems must facilitate the user access. Since both cannot happen at once, the application scenario then moves between two extremes: the highest-security (very few Type I errors) and the highest-convenience (very few Type II errors). Changing decision policies towards security, associated thresholds move from the extreme-convenience to the extreme-security and vice versa.

In order to map scores as system outputs of different recognition systems into the same calibrated score space, obtaining optimal calibration is the important step. Afterwards, relations between calibrated thresholds, i.e., formalized decision policies, are addressable. Thereby, relationships between decision policies become mathematically expressible and requirement assessment becomes automatically provable. As an LLR threshold η is logarithmic, an equally important movement of the threshold to higher security applications and to higher convenience applications will have the same absolute value (although it will have opposite signs). For instance, multiplying the cost-ratio $\frac{c_B}{c_A}$ from 10 to 100, i.e., times 10 towards higher-security applications, will mean an increase in η of $\log(10) = 2.3$; and an equivalent multiplication of $\frac{c_B}{c_A}$ from 100 to 10, i.e., times 10 towards high-convenient applications, will add -2.3 to η . The same applies to the multiplication of the ratio of prior probabilities. Thus, a calibrated domain is desirable in order to measure how equivalent changes in the application (changes in η) affect the system. In a calibrated score scale, moving the Bayes logarithmic threshold η a given quantity (e.g., by ± 2.3) implies the same for all score distributions since the scores are well-calibrated, i.e., interpretable, as the meaning of each LLR value is directly linked to the meaning of their corresponding LLR threshold η . If both values are the same, the score reports on the ev-

²³ Scores are treated as LLR_s and not LR_s due to mathematical convenience and in order to assess log-odds.

idence in the trade-off between class proportions and the threshold informs on the least required trade-off between prior and cost odds associated with the same classes.

In the following segment, the error rate centered perspective of the biometrics standardization community is introduced. Gaps are pointed out regarding the BDF, employed within the speaker recognition community, and the difference in implications of both perspectives are pointed out by two examples. Afterwards, the BDF is explained in detail and depending perspectives on performance assessment are outlined. An overview on automated speaker recognition systems employing the BDF is provided thereafter, depicting the state-of-the-art speaker recognition system that constitutes this dissertation's baseline system.

2.2 BIOMETRIC SYSTEMS IN ISO/IEC STANDARDIZATION

When employing biometric systems, harmonization between *biometric system vendors* (developing biometric systems) is crucial in order to sustain high performance, low error rates as well as low decision risks, good *sample quality* and security to *biometric system owners, operators, users (of a biometric system)*, and *biometric system end-users* of biometric systems. The ISO/IEC JTC 1/SC 37 *biometrics*²⁴ committee develops standards with relevance to, e.g., passports, border control, and biometrics in banking and payments.²⁵ SC 37 develops standards within different working groups²⁶ (WGs):

- **WG 1:** *harmonized biometric vocabulary*
- **WG 2:** *biometric technical interfaces*
- **WG 3:** *biometric data interchange formats*
- **WG 4:** *technical implementation of biometric systems*
- **WG 5:** *biometric testing and reporting*
- **WG 6:** *cross-jurisdictional and societal aspects of biometrics*

Some terminology of the speaker recognition community is employed.²⁷ However, as the scope of the *harmonized biometric vocabulary*

²⁴ ISO/IEC JTC 1 is the joint technical committee (JTC) on *information technology* between the *International Organization for Standardization* (ISO) and the *International Electrotechnical Commission* (IEC). JTC 1 is organized in subcommittees (SCs).

²⁵ Parts of this section are based on a collaborative work with Christoph Busch [84].

²⁶ See <https://www.iso.org/committee/313770.html>, accessed: 2017-10-12.

²⁷ The term *speaker recognition* is the example par excellence for conflicting terminology definitions. The community refers to itself as *speaker recognition*, whereas the SC 37 compliant term would be *voice biometrics* instead. In the SC 37 harmonized vocabulary, *recognition* is the umbrella term for verification and identification tasks. However, in the speaker recognition community, *speaker recognition* can also refer to

[1] is placed regarding black (and gray) box testing,²⁸ this dissertation partially differs in some terms. Exemplarily, the term *sample quality* is specified within biometric standardization as its use within standardization diverges from its use in natural language. Otherwise, the *Concise Oxford English Dictionary* (COED) [86] is sufficient for defining terminology. In biometric standardization, the quality of a sample is explicitly linked to degrading biometric recognition performance. In the COED, one of its definitions is specific to voice data: *distinguishing characteristic or characteristics of a speech sound* [86]. The purpose of this research is to sustain good performance levels regardless of the quality. This is only possible if quality is linked to signal properties instead of performance (as is the case in biometric standardization). The COED definition of *quality* as *the degree of excellence of something* leaves ambiguity to the interpretation of *excellence*. For standardizing the evaluation of biometric systems for a fixed dataset, linking *excellence* with performance is sufficient. For researching on how to improve performance on non-fixed datasets, however, this link is insufficient. This dissertation's approach of quantifying excellence aims at examining changes within signal processing of a recognition system; quality changes of a voice sample are linked to changes in its representation during speech signal processing.

This section depicts relevant aspects of WGs 1 and 5.

2.2.1 Generalized System Design

The design of biometric systems is standardized²⁹ in terms of a general subsystem composition [25, 87]. Fig. 2.2 (as of [87]) illustrates the subsystem composition. The subsystems are:

- **Data capture.** *Biometric capture subjects* present their biological or behavioral characteristics, which are captured by biometric sensors (potentially assembled in a capture device), e.g., microphones, resulting in a biometric sample, e.g., a digital audio file.

voice verification as the identification task is usually researched in terms of *language recognition* (technology-pushing labs working on speaker recognition mostly work on language recognition simultaneously). The BDF provides fairly enough freedom to generalize from one task to another. This choice is made for the purpose of the *visibility* of this dissertation in its major research field, the speaker recognition community.

²⁸ The main purpose of ISO/IEC 2382-37 [1] is the harmonization of documents developed within the ISO/IEC JTC 1/SC 37 *biometrics* committee [85], in particular towards the ISO/IEC 19795-1 project [25]. As [25] standardizes the performance and testing reporting of biometric systems, the communication between vendors, operators, and owners is targeted. For the sake of tractability and in order to reach consensus on performance measures and visualizations, abstract perspectives on biometric systems are solely considered, particularly black (and gray) box testing.

²⁹ See standing document (SD) 11 of ISO/IEC JTC 1/SC 37 [87], <http://isotc.iso.org/livelink/livelink?func=ll&objId=9626779&objAction=Open>, accessed: 2017-10-12.

- **Signal processing.** Samples are segmented into relevant and non-relevant regions of interest, comprising *speech* and *non-speech*, such that biometric features can be extracted. Quality control assures high system performance by rejecting low-quality samples, assuming that new samples can be recaptured.³⁰ High-quality features representing *enrolment* samples are utilized for the creation of biometric references. Probe samples that are assured to be of high-quality as well are utilized in the comparison subsystem.
- **Data storage.** Databases comprise enrolled biometric references. For recognition comparisons, stored references are requested and loaded from the database. During *verification*, a single biometric identity is claimed and the depending reference is loaded. During *identification*, multiple references are loaded.
- **Comparison.** Biometric reference and probe features are compared, reporting their similarity or dissimilarity as scores, whereby *verifications* are 1 : 1 reference–probe comparisons, and *identifications* compare a set of references against a probe, i.e., 1 : n comparisons.
- **Decision.** In a two-stage process, verification or identification outcomes are determined. First, scores are compared to thresholds, resulting in preliminary decisions or candidate lists. Second, decision policies are employed on these preliminaries from which the final recognition outcome results.

For the purpose of putting this dissertation into the context of biometrics standardization, that is to generalized biometric systems, the focus is solely³¹ placed on *verification* as a recognition task, particularly in similarity scores resulting from the comparison module.

Example: Generalized Biometric Systems in this Dissertation

In this dissertation, data capture, data storage, signal processing, and comparison subsystems are assumed to be fixed, whereby the data flow is readable. Moreover, the decision subsystem (likewise the jurisprudential decision maker) is segregated, such that the score output of the comparison subsystem needs to be well-calibrated^a in order to satisfy different decision policies that are un-

³⁰ Recaptures can be sustained in active capture scenarios, such as border control or mobile payments. In forensic biometrics, however, crime scene specimen cannot be recaptured, leading forensic experts to assessing the benefit of examining a specimen's sample quality in order to continue with analytic evidence reporting or refusing the specimen.

³¹ Notably, the employed comparator modules are also capable to 1 : 1, 1 : n and m : n recognition tasks, such that measures and methods proposed in this thesis are expected to be generalizable. For the purpose of keeping the narrative rather digestible, speaker recognition is addressed in terms of *verification*.

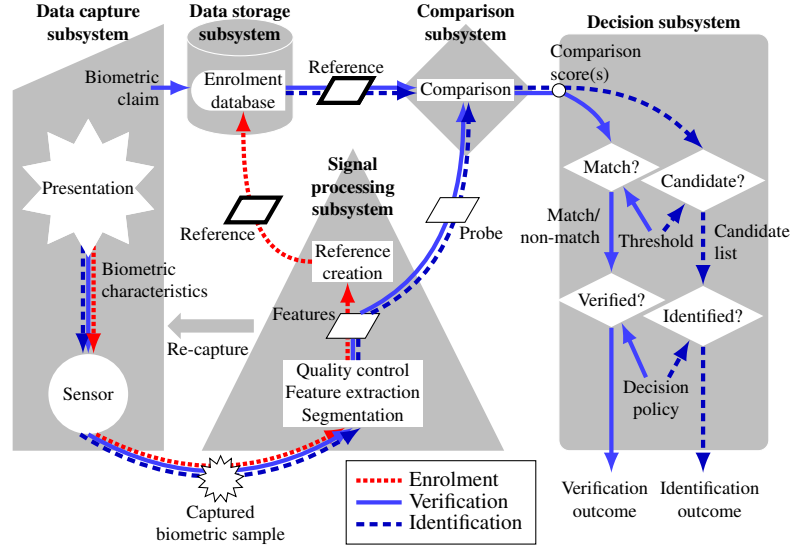


Figure 2.2: General design of a biometric system; source and shape semantics: ISO/IEC JTC 1/SC 37 SD 11 [87].

known at the time of system development or when operational conditions change in unconstrained environments.

Regarding forensic practitioners, the evidence of biometric voice data cannot be (re-) captured as in controlled environments and hence needs to be examined *asis* or be rejected for case work. Thus, *sample quality* is addressed in this dissertation as a quantification in terms of an estimate about a condition's characteristic for the purpose of adaptively normalizing and calibrating comparison scores to improve decision making for decision policies unknown to system development. In contrast to the *quality control* process that could issue a recapture of biometric samples, as depicted in Fig. 2.2, quality information estimated during feature extraction could aid in recalibrating *log-likelihood ratio (LLR)* scores (or thresholds).

In order to analyze the impact of varying environmental conditions that are directly comparable, the signal processing subsystem is investigated distinctively in terms of *acoustic* and *biometric* feature extraction. Intermediate features situated in between are analyzed in order to estimate quality impacts on the non-biometric signal processing. As an extension to the generalized design, this dissertation proposes to exploit the *quality control* module under simulated operational condition changes, such that the behavior of system decisions is adaptively calibrated using quality estimates. Preserving scores as the *weight of evidence* (LLRs) sustains good decision making in forensic and commercial application use cases.

^a Notably, published SC 37 standards (as of 2019) do not consider score calibration.

2.2.2 Performance and Testing Reporting Framework

In order to fairly compare biometric systems, ISO/IEC 19795-1 [25] defines a framework for performance testing and reporting. The framework differentiates fundamental performance metrics on error rates at different stages of a biometric system, in particular:

- **Enrolment and acquisition.** The proportion of uncaptured biometric characteristic presentations is the *failure-to-acquire rate (FTA)*, whereas the proportion of captured enrolment samples but of insufficient *biometric sample quality* is the *failure-to-enrol rate (FTE)*.
- **Comparison.** As algorithm performance, the Type I error rate is the *false match rate (FMR)* and the Type II error rate is the *false non-match rate (FNMR)*.
- **Transaction.** Combining the FTA and both algorithm error rates in technology evaluations, the Type I error rate is the *false accept rate (FAR)* and the Type II error rate is the *false reject rate (FRR)*.
- **Technology, scenario, and operational evaluation.** Combining the FTE and the transaction error rates, the Type I error rate is the *generalized false accept rate (GFAR)* and the Type II error rate is the *generalized false reject rate (GFRR)* in scenario or operational evaluation. Thereby, the scenario evaluation concerns end-to-end prototypes, and the operational evaluation is conducted for one specific population and application combination.

In this dissertation, the FMR and the FNMR metrics are solely utilized as the conducted research subdues technology evaluations. The FMR is the *proportion of completed non-mated comparison trials in which the non-mated probe and reference are falsely declared “match”* [25, 2019 CD1 revision]. The FNMR is the *proportion of completed mated comparison trials in which mated probe and reference are falsely declared “non-match”* [25, 2019 CD1 revision].

FMR and FNMR trade-offs are visualized by so-called *detection error trade-off (DET)* plots [88], cf. Fig. 2.3. For the sake of easier tractability, scores stemming from mated or non-mated comparison trials are referred to as class \mathcal{A} and class \mathcal{B} scores. DETs are originally motivated by observations of the speaker recognition community that many (adaptive) score normalization techniques tend to Gaussianize the score distributions [89]. The motivation of DETs is to scale the Type I and Type II error rate axes in such a way that error trade-offs rather resemble straight lines than curves to aid the visualization for an easier separability of systems' performance. Therefore, the

quantile function of the $\mathcal{N}(0, 1)$ distribution is used, i.e., the inverse cumulative density function of the standard normal distribution. Thus, DETs are quantile-quantile (Q-Q) plots of error rates. For the purpose of illustrating capabilities and limitations of DET plots, Figs. 2.3a to 2.3c depict synthetic score distributions. Arbitrary system outputs are simulated over different score ranges. (These distributions serve as exemplary system outputs and are not intended to model some error rate properties.) Depending score sets $S_{2.3a}$, $S_{2.3b}$, $S_{2.3c}$ are sampled from normal (\mathcal{N}), Beta (Beta), chi-square (χ^2) and uniform (\mathcal{U}) distributions: if score distributions follow Gaussian distributions, their trade-off resembles a straight line in this Q-Q plot. However, if a straight line resembles, the underlying score distributions are not necessarily Gaussians as seen for the score set $S_{2.3b}$, sampled from χ^2 and Beta distributions. Straight lines occur for other score distributions than the normal distribution as well in the DET space.

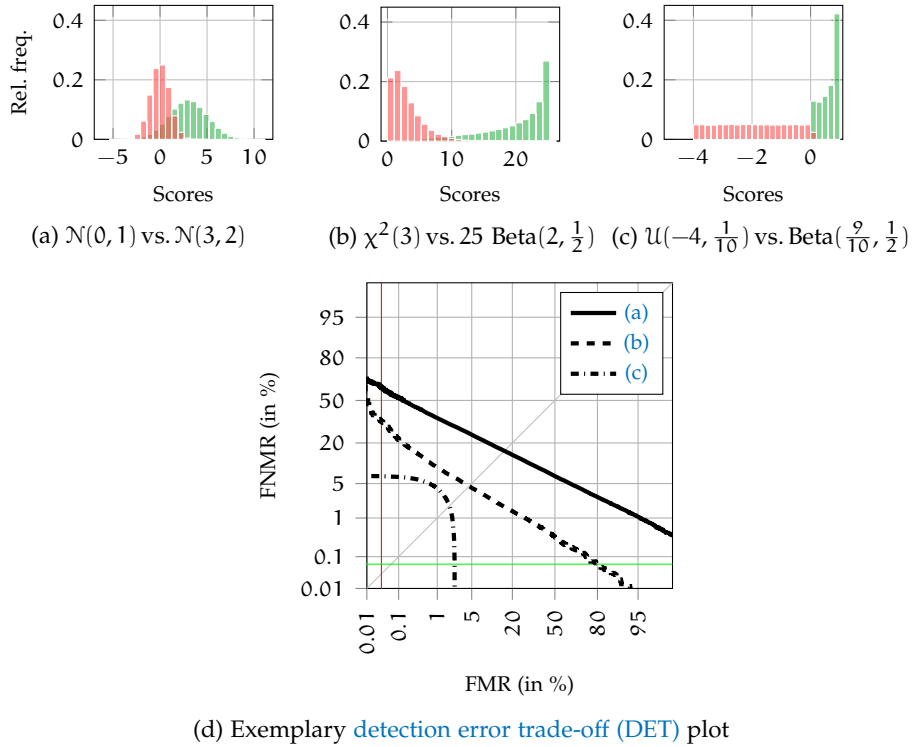


Figure 2.3: Examples of DETs for synthetic score distributions: histograms (10^5 class \mathcal{B} versus 5×10^4 \mathcal{A} score sets) (a),(b),(c) with class \mathcal{B} scores (red), class \mathcal{A} scores (green); (d) DETs of (a) (solid), (b) (dashed), and (c) (dash-dotted) with **rule of 30** indications for **FMR** (red) and **FNMR** (green).

In contrast to **receiver operating characteristics (ROCs)** [90], DET plots can depict neither 0% nor 100% due to their underlying quantile function, which considers a $(-\infty, +\infty)$ space. Therefore, the major advantage of DET plots as described in [88] is that they are visually easier distinguishable by the human eye when score distributions are

Gaussians. The *quantile* function of $\mathcal{N}(0,1)$ is also known as the *probit* function, defined by the inverse error function erf^{-1} as:

$$\text{probit}(x) \equiv \sqrt{2} \text{erf}^{-1}(2x - 1). \quad (2.1)$$

For the purpose of denoting upper and lower bounds to observed error rates, the standardized performance testing and reporting framework [25] depicts Frequentist uncertainty estimates in terms of confidence intervals, particularly two rules of thumb: the [rule of 3](#) [91] and the [rule of 30](#) [92]. Furthermore, [zero-effort impostors](#) are considered, whereas attacks targeting a specific biometric reference are not. Attackers might produce elaborated [presentation attack instruments](#) (PAIs) to yield higher similarity scores against the biometric reference of a specific target subject in a comparison trial.

2.2.3 Presentation Attack Detection: Testing and Reporting

[Presentation attacks](#) comprise *artificial, human or other natural* PAIs [43], which are presented at sensor level to the biometric system but may also consider modified biometric samples in research evaluations [93]. Artificial PAIs are comprised [presentation artefacts](#), such as in audio replay attacks, speech synthesis or voice conversion. Fig. 2.4 illustrates attack points to a biometric system and depicts the scope of presentation attacks according to ISO/IEC 30107-1 [43].

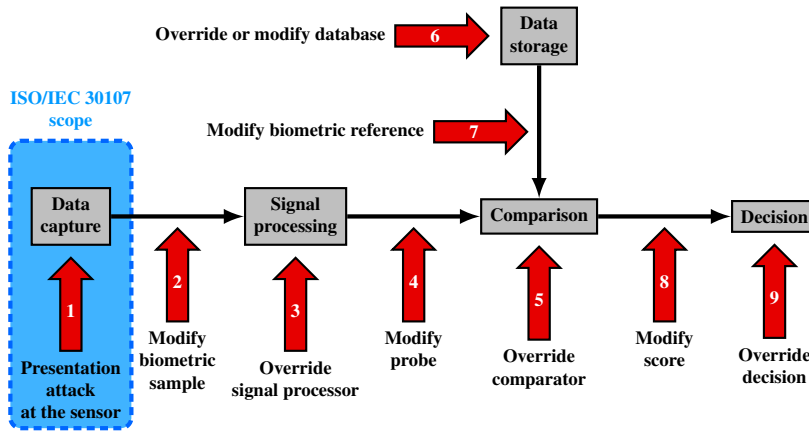


Figure 2.4: Overview on attack points in biometric systems with depicted scope of ISO/IEC 30107-1; source and shape semantics: [43].

For detecting presentation attacks, [presentation attack detection](#) (PAD) is considered an additional subsystem to the biometric system [43], see Fig. 2.5: the PAD subsystem might influence the signal processing or the comparison subsystems. However, [43] is not restricting the system design as such. In [45, 94], system designs consider either score fusion of comparison scores with PAD subsystem scores or a unified classifier jointly conducting biometric comparison and PAD recognition tasks.

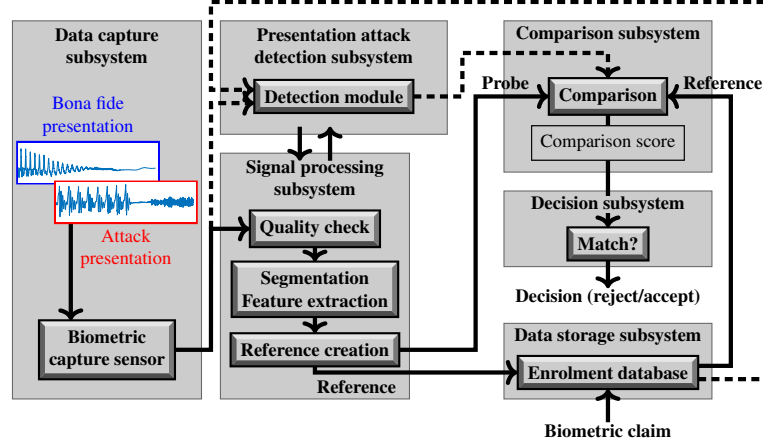


Figure 2.5: General biometric system composition with PAD; source and shape semantics: ISO/IEC 30107-1 [43].

For operational evaluations, ISO/IEC 30107-3 [93] proposes the performance assessment utilizing error rates for the PAD subsystem, inspired by ISO/IEC 19795-1 [25]. For scenario and technology testing, the biometric standard family on PAD proposes to employ synthetic PAIs for offline testing. Independent of whether presentation attacks are conducted by *impostors* (subverting a system to be recognized as another identity, by Type I errors) or by *identity concealers* (subverting a system to remain unrecognized, by Type II errors), Type I and Type II error rates are represented by the *attack presentation classification error rate (APCER)* and the *bona fide presentation classification error rate (BPCER)*. The APCER is the *proportion of attack presentations using the same PAIS incorrectly classified as bona fide presentations in a specific scenario* [93]. The BPCER is the *proportion of bona fide presentations incorrectly classified as presentation attacks in a specific scenario* [93]. For the purpose of targeting operating points enforcing security at a defined level, [93] recommends (but does not require) to report the BPCER at a 5% APCER (BPCER₂₀). Moreover, figures of merit based on weighted error rates are explicitly *deprecated* in this standard—a result from discussions on the half total error rate. Information theory motivated measures were not considered as the depreciation was agreed upon. The APCER and BPCER metrics are relevant to this dissertation’s emphasis on PAD. Other PAD related metrics consider non-responses, duration efficiency, the data capture subsystem, and full system evaluation.

The performance and testing reporting framework as such, however, solely depicts a Frequentist’s perspective on performance assessment, whereas this dissertation aims at the benefit of a system’s contribution to decision making, such as decision risk assessment employing the *BDF*. In contrast to the biometric standardization community, the ongoing PAD research challenge of the speaker recognition community, the ASVspoof challenge, employs a tandem decision risk

[52] as a figure of merit in its 2019 edition [48], i.e., a weighted error rate motivated from information theory resembling the expected risk of employing a speaker recognition system as well as a PAD subsystem. Both communities disconnect: in ISO/IEC 30107-3 [93], any form of *weighted error rates* are deprecated, such as the ASVspoof 2019 primary figure of merit. This discrepancy illustrates that more harmonization work is necessary to bring both (currently rather disjoint) wings of the larger biometric community closer together.

2.3 GAP ANALYSIS ISO/IEC 19795-1:2006

The ISO/IEC 19795-1:2006 [25] standard is one of the most relevant³² projects of the ISO/IEC JTC 1/SC 37 committee as it is focussed on the principles and framework of biometric performance testing and reporting.³³ Its major benefit is a unified reporting of biometric recognition performance *as is*, agreed under international consensus from industrial, governmental, and academic sites. However, the 2006 version of this standardization project remains unchanged (confirmed in 2011 and in 2016), such that more recent methodologies in performance assessment have been unconsidered for over a decade.

In brief, the gap of ISO/IEC 19795-1 [25] to the state-of-the-art performance evaluation (in speaker recognition) is due to the fact that the standard has not been changed for over 15 years (when it was initially drafted). The standard solely considers evaluation results in terms of error rates *as is* rather than their *predictive power*. Thereby, empirical uncertainty is just addressed in terms of *the trust in repetitions* (confidence in a sampling comprising true values) but not in terms of *the trustworthiness of forecasts* (credibility in the inference among multiple uncertain decisions). Considering the *Bayes theorem* not only in system development (machine learning) but also in system evaluation, the following gaps of the ISO/IEC 19795-1:2006 standard [25] to the speaker recognition community are identified:

³² Historically (from the start of the SC 37 committee onwards), the early 19795-1 project aimed at the evaluation performance harmonization of multiple vendor algorithms. Beside other projects from the 19795 family (required to implement their part 1 framework), the standardization of biometric application interfaces (province of WG 2, not required to implement 19795-1) builds upon the design of a generalized biometric system (province of WG 5), cf. Fig. 2.2. For the *harmonized* communication between standardization projects being discussed in different WGs, two standing documents (SDs) were established: SD 2 on the *harmonized biometric vocabulary* [95], regularly updating the ISO/IEC standard 2382-37 [1]; and SD 11 [87] on the *general biometric system*, summarizing the outline of the 19795-1 standard [25]. Based on the interface specifications (province of WG 2), technical implementations are standardized (province of WG 4), independent of data interchange format specifications (province of WG 3). In order to pursue a *harmonized* terminology, SD 2 is an ongoing project (province of WG 1) to which all other WGs contribute (also WG 6 on cross-jurisprudential and societal aspects). In other words, the outline and assumptions of the 19795-1 project [25] are fundamental to many SC 37 projects.

³³ Parts of this section are based on a collaborative work with Christoph Busch [84].

- **Proper scoring rules.** The applicability of a biometric system to decision making processes is not addressed. As testing reports are likely to be employed for performance prediction of future decision making, the prediction accuracy assessment of such reports is of utmost relevance to avoid *misleading decisions* that could be made when performance reports solely report error rates *as is* regardless of their impact to decision making.
- **Identity inference for decision making.** Bayesian (identity) inference is crucial for the purpose of making good decisions on average and, consequently, in order to assess the depending Bayes (decision) risk as a performance metric. The standard addresses a traditional *Frequentist* perspective by limiting its scope to the exclusive reporting of error proportions, sparing any *Bayesian* perspective on performance assessment. Consequently, Bayesian perspectives on performance evaluation appear not to comply with the standard (as non-conformant) as depending methods remain unmentioned.

The current standard addresses *confidence* intervals of error rates rather than *credible* intervals: *confidence* intervals concern a large number of repetitions, e.g., 95% of (entire) experiment samplings (repetitions) confirm an *error rate interval* and 5% of the samplings oppose this error rate interval, such as 5 out of 100 conducted experiments. By contrast, *credible* intervals concern the *uncertainty of an error rate estimate*. For a fixed 95% *credibility*, the true error rate is believed to be in a narrower interval around the observation when larger data amounts are used. In that case, the error rate estimation is believed to be more precise. When using smaller data amounts, the true error rate is believed to be in a wider interval around the observation. The error rate estimation is then believed to be more uncertain/less precise. Even if the debate on Frequentist versus Bayesian is a rather philosophical one, it is possible to state that, in an experimental observation, the Frequentist's perspective on the current standard can only provide answers to where a true error rate is indirectly, whereas a Bayesian perspective provides answers directly.

- **Information theory.** This standard focuses on particular application requirements for systems under test, with error rates serving as a proxy for an information theoretic perspective on performance. However, under unknown requirements prior to the employment of a system, such as during research and development of multi-application systems or for forensic scenarios, an abstraction of performance is necessary, especially when decision cost matrices remain unspecified (e.g., in a forensic witness report). Therefore, threshold values need to be addressed

in a more formalized way. By associating their value corresponding to the proportion of competing classes within the feature space, the divergence between score distributions of each class is addressable in a formal manner. The formalization of decision making allows to examine the *information gain* from employing a system for an (identity) inference process under different decision requirement assumptions.

These gaps represent two fundamentally different approaches on *denoting decision thresholds*: the standard assesses proportions of errors initially in order to derive the corresponding operating point, a threshold value. Under the Bayesian paradigm, in contrast, threshold values are denoted before performance evaluation. However, for the premise of testing solely *one* application scenario with *one* report, i.e., no cross-application testing, the 19795-1 standard [25] is well applicable, such as in research and development *solely dedicated* to automated border control (between two specific airports). The outlined gaps need to be considered for research and development of cross-application systems, e.g., for smart home, online banking, or forensic applications. In the following, the gaps are addressed in more detail. The following sections briefly discuss proper scoring rules, Frequentist versus Bayesian philosophies and their implications towards the [rule of 3](#) and [rule of 30](#), parts of the informative (non-normative) annex of ISO/IEC 19795-1 [25]. Afterwards, identity inference for decision making is addressed in more detail, illustrating perspectives on performance measures emerging (inter alia) from the speaker recognition community. Finally, the state-of-the-art in speaker recognition and the baseline system used in this dissertation are outlined.

2.3.1 Proper Scoring Rules: in Brief

In *decision theory*, the accuracy of (probabilistic) predictions is addressed by *proper scoring rules* [78, 96–98], with Brier [96] being the most prominent paper. In this paper, the *goodness* of weather forecasts is addressed, proposing to relate the *forecast probability* with the *relative occurrence probability* (error proportions). It states that *arbitrary scores may not be the most useful forecast*. The metric *Brier score* is the mean square error between prediction probability and the actual outcome (favoring a proposition), i.e., between a system's (probabilistic) score and the probability of that proposition to occur. In the literature, various other scoring rules are proposed. The *goodness* of automatic pattern recognizers is evaluated by their *application-independent Bayes risk*: scores are recognition system outputs that need to reflect the class distribution within the feature space (in a formalized manner). In other words, scores need to encode Type I and Type II error proportions when making binary decisions, independent of the classes' prior probabilities that are conveyed within an evaluation dataset. That is,

to produce **log-likelihood ratio (LLR)** scores, where the depending expected Bayes risk is referred to as the *cost of LLRs*, namely C_{llr} [26, 28]. In the feature space, LLR *scores* reflect the class distribution regarding the competing propositions. In the decision space, LLR *thresholds* reflect the ratio of erroneous decisions made regarding these propositions (considering the prior probabilities of classes in a dataset). LLR scores are the most useful scores for binary decision making. Motivated by proper scoring rules, the preservation of LLR score properties is essential to sustain *better decision making on average*.

Proper scoring rules may be defined in two equivalent ways [99]:

- Let P and Q be probability distributions predicting a class $u \in \mathcal{U}$. Proper scoring rules are real-valued functions $S(Q, u)$, where the expected value³⁴ w.r.t. P is minimized at $P = Q$:

$$\langle S(P, u) \rangle_{u \sim P} \leq \langle S(Q, u) \rangle_{u \sim P}. \quad (2.2)$$

If the minimum is unique, the rule is referred to as a **strictly proper scoring rule**.

- Let Q be a probability distribution to predict a class $u \in \mathcal{U}$. Let $d \in \mathcal{D}$ be a *decision* based on Q , while u is (still) unknown. Let $C(d, u)$ be a real-valued cost function, quantifying the *goodness* of decision d , after the class of u becomes known. Let $d_Q^* \in \mathcal{D}$ be a minimum-expected-cost decision with:

$$\langle C(d_Q^*, u) \rangle_{u \sim Q} \leq \langle C(d, u) \rangle_{u \sim Q} \quad \forall d \in \mathcal{D}, \quad (2.3)$$

then $S(Q, u) = C(d_Q^*, u)$ is a proper scoring rule.

Both definitions imply another definition; let $\mathcal{D} = \mathcal{U}$: decisions d can be interpreted as predictions for u [99].

2.3.2 Frequentist and Bayesian: in a Nutshell

Philosophically [100], Frequentist and Bayesian approaches elaborate on concepts of uncertainty. Frequentists describe uncertainty *due to randomness* (aleatory uncertainty) but cannot describe the uncertainty *due to lack of knowledge* (epistemic uncertainty) using probabilities. In contrast, Bayesians utilize probabilities to quantify any kind of uncertainty: a Bayesian proposition's probability represents a *degree of belief* in the truth of that proposition (randomness in a Frequentist's perspective).³⁵ In other words, a Bayesian *credible interval* addresses the chance of a *proposition* being true, while the Frequentist *confidence interval* addresses the relative amount of experiments, which will contain the true value, denoted as p-value like, e.g., $p = 0.95$. The heart

³⁴ The expectation operator is denoted by $\langle \cdot \rangle$.

³⁵ Again, the term *proposition* is preferred over the terms *event* or *hypothesis*, which imply randomness of the outcome.

of the debate is on *whether something is unknown or unknowable, whether its uncertainty is due to fundamentally unpredictable randomness or to potentially resolvable lack of knowledge* [100].

By consequence, this debate addresses the inference of parameter values of a recognition mode: Frequentist approaches can exclusively infer statements about parameters indirectly, whereas Bayesian approaches can make direct statements that are unambiguous towards the parameters to be learned. Thereby, from a Frequentist perspective, error rates represent the discrimination power of a learned model, whereas in a Bayesian perspective, performance is quantified directly regarding both the discrimination power and the prediction power. The performance evaluation of biometric recognition needs to be addressed regarding proper scoring rules in order to inform commercial and forensic system *biometric system owners* about the usefulness of employing a system for their decision making requirements. Regarding biometric recognition, this dissertation solely emphasizes the verification task, i.e., LLRs in 1 : 1 comparisons with the C_{llr} proper scoring rule ([28] provides a multi-class generalization of C_{llr}).

From a Frequentist perspective, thresholds are indirectly defined by error rates: an error rate constraint is defined before evaluation, a system's threshold value achieving this constraint is derived *after* all comparison scores are computed. Thereby, the decision policy is not necessarily considered right away. For instance, this is the case in biometric standardization [25, 101], where a *match/non-match* decision is made utilizing an error rate based threshold. The *verification outcome*, however, may employ further decision policies, see section 2.2, leading to a formal way of denoting operating points (*Bayes thresholds*). In the following section, implications of either beformentioned philosophies on performance assessment are discussed and compared by means of two examples. Both examples briefly address how error rates are principally dealt with. Implications and benefits of the Bayesian perspective are directly compared to well-established tools based on a Frequentist perspective. After these examples, the Bayesian perspective is introduced in depth.

2.3.3 Thresholds According to Error Rates (Frequentist)

An error occurs by comparing a score S to a threshold t :

$$\begin{array}{c} \text{accept} \\ t \leq S. \\ \text{reject} \end{array} \quad (2.4)$$

A Type I error occurs if a class \mathcal{B} score is greater or equal to t ; and a Type II error occurs if a class \mathcal{A} is lower than t . Looking at one error type at a time, each comparison is a trial of binary outcome, whereby a series of binary-outcome trials is known as a *Bernoulli experiment*. Conventionally, Bernoulli experiments are outlined for coin tosses.

This setup describes the toss of an unfair coin with the depending error rate resembling the fairness of that coin. When examining a score set instead of coin tosses, an error rate e is drawn as a Binomial experiment with the underlying Binomial distribution $\text{Bin}(n, p)$ with n i.i.d. (comparison) scores and the true error probability p and k being the number of errors:

$$\Pr(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}. \quad (2.5)$$

Thereby, the observed error rate e is a sampling of $\text{Bin}(n, p)$; the observation e is not necessarily equal to the truth p .

As the observed error rate is not the true error rate—after all, we are sampling—precise statements about recognition performance are of interest. Two rules of thumb, namely the [rule of 3](#) [25, 91] and the [rule of 30](#) [25, 92] are motivated from the Frequentist perspective but find their counterparts in the Bayesian perspective as well. Both denote bounds to the estimation of the true error probability p based on the observed error rate e . The rule of 30 addresses how much the true error rate is off from the estimate derived from the sampling and how many observed errors would be necessary to limit the uncertainty of error rate estimates. The rule of 3, however, addresses the observation of zero errors. In the case of observing perfect class discrimination, the lowest error rate that can be reported under a given confidence is estimated.

Example: Frequentist Implication I/II, the Rule of 3

When observing zero errors, a Frequentist questions the randomness of a sampling, an experimental setup in this case, as other i.i.d. samplings might result in higher observed error rates. The binomial distribution of Eq. (2.5) is parameterized with $k = 0$ (zero errors) and compared to a confidence interval c :

$$\begin{aligned} \Pr(0; n, p) &\geq 1 - c, \\ \binom{n}{0} p^0 (1-p)^{n-0} &\geq 1 - c, \\ (1-p)^n &\geq 1 - c, \\ n \log(1-p) &\geq \log(1-c). \end{aligned} \quad (2.6)$$

The term $\log(1-p)$ depends on the true error rate p , which is expanded by the Taylor/Maclaurin series (as $|p| < 1$):

$$\log(1-p) = -\sum_{k=1}^{\infty} \frac{p^k}{k} = -p - \frac{p^2}{2} - \frac{p^3}{3} - \dots. \quad (2.7)$$

For very small p values, the higher moments become negligible, such that: $\log(1-p) \approx -p$. The least upper bound to p is approximated by:

$$-np \approx \log(1-c), \quad \approx \frac{-\log(1-c)}{n}. \quad (2.8)$$

The true error is estimated depending on the number of i.i.d. scores. The rule of 3 states for a $c = 95\%$ confidence:

$$p \approx \frac{3}{n}; \quad (2.9)$$

the quantile value for $c = 0.95$ is approximated as $\log(0.05) \approx -3$.

The rule of 3 is a rule of thumb depending on the number of scores, presenting the lowest error rate for a 95% confidence level. When re-sampling an experimental setup, the prediction of p is the lowest true error rate in 95 of 100 cases *but not in 5 of 100 cases*, on average.

Example: Frequentist Implication II/II, the Rule of 30

When observing errors, a Frequentist questions the randomness of a sampling, as in other i.i.d. samplings the true error rate might be completely off from the observed error rate. In [25, 92], a relative band is hypothesized around the observed error rate e in terms of a scaling term α (the relative band is $(1 \pm \alpha)e$), wherein the true error rate could lie (or not). The difference between observation and truth $|e - p|$ compared with a relative band of the observation αe by a scaling factor α is expressed probabilistically: $\Pr(|e - p| \geq \alpha e)$. In order to simplify comparisons to a confidence interval c , this probability term can be expressed without the absolute value operator $|\cdot|$ as:

$$\Pr(|e - p| \geq \alpha e) = \Pr\left(e \geq \frac{p}{1 - \alpha}\right) + \Pr\left(e \leq \frac{p}{1 + \alpha}\right) \leq 1 - c. \quad (2.10)$$

Akin to the rule of 3, a binomial distribution is assumed. By assuming n is large, the binomial distribution is approximated by a normal distribution which is known as the *Wald test*: $\text{Bin}(n, p) \approx \mathcal{N}(np, np(1 - p))$. Thereby, the probability terms $\Pr(e \geq \frac{p}{1 - \alpha})$ and $\Pr(e \leq \frac{p}{1 + \alpha})$ are assumed to follow a normal distribution, for which p is standardized in the form of $\hat{p} \sim \mathcal{N}\left[p, \frac{p(1 - p)}{n}\right]$. For the probability $\Pr(e \geq \frac{p}{1 - \alpha})$, this standardization is carried out as:

$$\begin{aligned} & e \geq \frac{p}{1 - \alpha} \\ \Leftrightarrow & e - p \geq \frac{p}{1 - \alpha} - p = \frac{p - p + \alpha p}{1 - \alpha} = \frac{\alpha p}{1 - \alpha} \\ \Leftrightarrow & (e - p) \frac{\sqrt{n}}{\sqrt{p}\sqrt{1 - p}} \geq \frac{\alpha p}{1 - \alpha} \frac{\sqrt{n}}{\sqrt{p}\sqrt{1 - p}} \\ \Leftrightarrow & \frac{e - p}{\sqrt{\frac{p(1 - p)}{n}}} \geq \frac{\alpha}{1 - \alpha} \frac{\sqrt{p}\sqrt{p}\sqrt{n}}{\sqrt{p}\sqrt{1 - p}} \\ \Leftrightarrow & X \geq \frac{\alpha}{1 - \alpha} Y. \end{aligned} \quad (2.11)$$

The two random variables X, Y are introduced for the sake of easier tractability:

$$X = \frac{e - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}[0, 1], \quad \text{and} \quad Y = \sqrt{\frac{np}{1-p}}. \quad (2.12)$$

Indeed, e is centered by the mean value p and normalized by the standard deviation $\sqrt{\frac{p(1-p)}{n}}$. Similarly, the probability $\Pr(e \leq \frac{p}{1+a})$ is standardized:

$$\begin{aligned} e &\leq \frac{p}{1+a} \\ \Leftrightarrow e - p &\leq \frac{p - p - ap}{1+a} \\ \Leftrightarrow X &\leq \frac{-a}{1+a} Y, \end{aligned} \quad (2.13)$$

such that Eq. (2.10) is expressed by:

$$\Pr(|e - p| \geq a e) \approx \Pr\left(X \geq \frac{a}{1-a} Y\right) + \Pr\left(X \leq \frac{-a}{1+a} Y\right) \leq 1 - c. \quad (2.14)$$

Numerically, values for Y result by using the closed-form solution of the standard normal cumulative density function. The substituted Eq. 2.10 is solved to zero for the normally distributed random variable X :

$$0 = \Pr\left(X \geq \frac{a}{1-a} Y\right) + \Pr\left(X \leq \frac{-a}{1+a} Y\right) - (1 - c). \quad (2.15)$$

The solution will result as a term depending on Y , which encodes the terms p, n as: $Y^2 = n \frac{p}{1-p}$. The ratio $\frac{p}{1-p}$ is expressed by the Taylor series at the center 0 (assuming very small error rates):

$$\frac{p}{1-p} = \sum_{k=1}^{\infty} p^k = p + p^2 + p^3 + \dots, \quad (2.16)$$

where for very small p values, higher moments become negligible ($\frac{p}{1-p} \approx p$), such that: $Y^2 \approx np$; Y^2 approximates the number of true errors np . Parameterizing $a = 30\%$, $c = 90\%$ [25, 92] (30% relative band to the observation with a two-sided test; each side is tested with a 5% confidence interval), the number of errors approximates $Y^2 \approx 30$.

The rule of 30 addresses the minimum amount of errors necessary, such that the true error rate p lies within a relative band of $\pm a$ to the observed error rate e at a confidence interval c . In other words, for the shown parameterization, on average, at least 30 errors need to be observed, such that in 90 out of 100 cases, the true error rate p lies within a $\pm 30\%$ band of the observed error rate e ; *but not in 10*

out of 100 cases. In 5 out of 100 cases, the true error rate is below 0.7 e, and in another 5 out of 100 cases, the true error rate is above 1.3 e (values are limited to the [0, 1] interval for probabilities). Eventually, the rule of 30 assumes large datasets, such that the binomial distribution is approximated by a Gaussian distribution; and very low error rates, such that Y^2 approximates the number of true errors. On high error rates, the term $\frac{n \cdot p}{1-p}$ (number of true errors by the true success rate) remains rather intractable. Thus, the rule of 30 is inapplicable to datasets of limited data and to observed error rates that are higher than very low.

2.3.4 Thresholds before Evaluation (Bayesian)

This section illustrates how the outlines of the rule of 3 and the rule of 30 change when moving from a *Frequentist* to a *Bayesian* perspective. The uncertainty of the observed error rates is modeled by so-called *uninformative priors*, i.e., priors of *maximum entropy*. As a scalar, the value of the uninformative prior is 0.5 (a half or 50%). As a random variable (a prior distribution), different distributional assumptions can be made [64]. From a Bayesian perspective, a prior belief is modeled either by a scalar probability or by a probability distribution. The prior belief is updated by an observation to the posterior belief. If the distribution of posterior and prior belief is of the same probability distribution family, both distributions are called *conjugate*. Conjugate priors give a closed-form solution for the posterior (numerical convenience).

Example: Bayesian Implication I/II, the Rule of 3

The Bayesian rule of 3^a is described in [91]. The true error rate p given an observation Y is stated in terms of Bayes' theorem as [102]:

$$\Pr(p | Y) = \Pr(Y | p) \frac{\Pr(p)}{\Pr(Y)}, \quad (2.17)$$

where the posterior follows a Bernoulli distribution for error rates. The conjugate prior of the Bernoulli distribution is the Beta distribution. Parameters of the prior distribution are assumed as Beta(1, b) with the variable b , where the parameter 1 is chosen in order to provide the prior with a local maximum away from zero that, without additional information, cannot be justified [91]. Thereby, the corresponding Bayesian extension of Eq. (2.8) is approximated by a Taylor expansion with the least upper bound regarding the number of scores n [91]:

$$p \approx \frac{-\log(1 - c)}{n + b} \approx \frac{3}{n + b}. \quad (2.18)$$

The uninformative uniform prior is addressed by denoting $b = 1$ [91], yielding for a $c = 95\%$ credible interval:

$$p \approx \frac{3}{n+1}. \quad (2.19)$$

^a Notably, the ISO/IEC standard 19795-1 [25] depicts the Frequentist rule of 3 as of [91], while the Bayesian counterpart is spared.

A slight modification of the Frequentist rule of 3 results. With a credibility of 95%, the lowest error rate to be predicted is $\frac{3}{n+1}$ when observing zero errors on n scores.

Revisiting the rule of 30, which is rather suitable for large scale tests, the inherent assumptions of large n values might not hold true for small datasets. A Bayesian alternative for limited data scenarios is encouraged regarding the Type I and Type II error trade-off assessment. Error rates p might not be assumed small, either, if they are, for example, in the **equal error rate (EER)** region. On targeting challenging data, however, EERs might be fairly large as well. This raises the following question: given an observed error rate e and the number of i.i.d. scores n , which lower and upper bounds for the true error rate p to be in this relative band can be denoted on a 90% credible interval?

Example: Bayesian Implication II/II, the Rule of 30

Targeting an objective prior, an uninformative prior distribution can be used for a parametric space γ , such as Jeffrey's prior $p(\gamma)$ [103], which is related to the Fisher information matrix $\mathcal{J}(\gamma)$ as:

$$p(\gamma) \propto \sqrt{\det(\mathcal{J}(\gamma))}. \quad (2.20)$$

Regarding Bernoulli trials, Jeffrey's prior results in the Beta distribution: $\text{Beta}(\alpha = 0.5, \beta = 0.5)$, where α, β represent successes and failures. Given a score set of n scores, the observed error rate e is defined by the number of observed errors x with $e = \frac{x}{n}$, such that the posterior distribution for p is derived as:

$$p \sim \text{Beta}(x + 0.5, n - x + 0.5). \quad (2.21)$$

Inspired by the rule of 30, a two-sided 90% credible interval is examined, depicting the 5th-quantile^a of Eq. (2.21) as an estimate for the lower bound, and depicting the 95th-quantile of Eq. (2.21) as an estimate for the upper bound.^b

^a Example on $n = 90, x = 3, e \approx 3.3\%$ providing $p_{\text{lower}}^{\text{Bayes}} \approx 1.2\%$, see: <https://www.wolframalpha.com/input/?i=5%25+quantile+betadistribution%5Bx%2B.5,+n-x%2B.5%5D+with+n%3D+90+and+x%3D+3>

^b Example on $n = 90, x = 3, e \approx 3.3\%$ providing $p_{\text{upper}}^{\text{Bayes}} \approx 7.6\%$, see: <https://www.wolframalpha.com/input/?i=95%25+quantile+betadistribution%5Bx%2B.5,+n-x%2B.5%5D+with+n%3D+90+and+x%3D+3>

Tab. 2.1 provides a comparison among different database sizes on a fixed observed error rate between the Frequentist's rule of 30 to its Bayesian alternative: as the database size increases, the performance estimate becomes more certain, and the upper and lower bounds converge to the single point performance estimate. While the Frequentist confidence interval remains fixed, the Bayesian credible interval is capable of accounting for the amount of data. Its major benefit lies in the evaluations of low data amounts, e.g., in the prototype development of mobile biometrics and [presentation attack detection \(PAD\)](#) evaluations, where comparatively limited data amounts can be collected. In order to narrow the rule of 30 bounds (despite lowering confidence or credible percentiles), the Frequentist approach requires lower error rates, whereas the Bayesian approach requires more data.

Table 2.1: Examples of Bayesian rule of 30 credible interval $[p_{lower}^{Bayes}, p_{upper}^{Bayes}]$ of error rates for $e \approx 3.33\%$ on different database sizes n , compared with the Frequentist rule of 30 confidence interval $[p_{lower}^{Freq.}, p_{upper}^{Freq.}]$ with $p_{lower}^{Freq.} = (1 - 30\%) e$ and $p_{upper}^{Freq.} = (1 + 30\%) e$, remaining fixed regardless of n .

n	9×10^1	9×10^2	9×10^3	9×10^4	9×10^5	9×10^6
x	3	3×10^1	3×10^2	3×10^3	3×10^4	3×10^5
$p_{upper}^{Freq.}$	4.33%	4.33%	4.33%	4.33%	4.33%	4.33%
e	3.33%	3.33%	3.33%	3.33%	3.33%	3.33%
$p_{lower}^{Freq.}$	2.33%	2.33%	2.33%	2.33%	2.33%	2.33%
p_{upper}^{Bayes}	7.62%	4.43%	3.66%	3.43%	3.36%	3.34%
e	3.33%	3.33%	3.33%	3.33%	3.33%	3.33%
p_{lower}^{Bayes}	1.21%	2.46%	3.03%	3.24%	3.30%	3.32%

The examples of the rule of 3 and of the rule of 30 illustrate implications when moving from a Frequentist to a Bayesian perspective. Taking a Frequentist perspective, recognition performance evaluation concerns the repeatability of observations. After that, decisions are made in an *informal* manner: requirements are denoted on error rates (not on threshold values), such that threshold values are parameterized *after* an evaluation is carried out. By contrast, Bayesians *formalize* all types of uncertainty, including prior probabilities and decision costs, such that threshold values are parameterized *before* any evaluation is carried out. Then, the scope of recognition performance evaluations is extended to address the *goodness* of a recognition system's predictions. The following section illustrates how the observed gaps of the ISO/IEC 19795-1:2006 standard can be bridged when moving from *informal* to *formal* decision making on the biometric identity inference, i.e., from a Frequentist to a Bayesian perspective.

2.4 BAYESIAN DECISION FRAMEWORK

The **Bayesian decision framework (BDF)** combines Bayesian inference and decision theory [56, 57], see Fig. 2.1. In this dissertation, the BDF formalizes the interaction of the biometric comparison subsystem and the decision subsystem. Motivated by Bayes' theorem, prior beliefs of competing propositions (class \mathcal{A} : true identity claim, class \mathcal{B} : false identity claim) are updated by observing evidence in terms of **likelihood ratio (LR)** comparison scores outputted by the biometric system, resulting in a posterior belief towards each proposition. In other words, the biometric system contributes information to biometric comparison as evidence to a Bayesian inference. In Bayesian inference, a probabilistic prediction of each proposition results.³⁶ By comparing the probabilistic predictions, the posterior belief, to a decision cost matrix that reflects the operational policy, decision making is formalized. The decision outcome favors the proposition yielding the highest incentive, i.e., the lowest decision risk cost.

In this dissertation, (biometric) recognition systems, cf. Fig. 2.2, ideally estimate **log-likelihood ratio (LLR)** scores (log-LRs), and LLR thresholds are formulated based on the prior and cost beliefs. The posterior belief, as an intermediate result of the decision inference chain, is substituted with the cost belief, resembling the formal requirements of a decision cost matrix. LLR thresholds are parameterized by additive impacts of the prior belief and the decision cost matrix [28]. Eventually, an LLR threshold value is the prior belief (additively) biased by the cost belief but also vice versa; its parameterization can as well be seen as the cost belief biased by the prior belief. Either parameterization of the prior belief or the cost belief can remain unknown, e.g., some smart home applications might run in single households or in shared flats without further specification (the prior belief is of *full uncertainty*). Whereas in the current forensic practice, decision cost matrices are undefined since decisions are made in the province of court (on the *offense level*), forensic practitioners report on evidence at the *source* and *activity levels* [20]. At these levels, no intel exists on the cost belief—the cost belief is of *full uncertainty*. If the prior belief remains unknown (at *maximum prior entropy*), the decision cost matrix fully parameterizes the LLR threshold, and the impact of the prior belief contributes a zero bias. If the cost belief remains unknown (at *maximum cost entropy*), the prior belief fully parameterizes the LLR threshold, and the impact of the cost belief contributes a zero bias.

Effectively, decision and comparison subsystems are decoupled [24]: *subjective* thresholds are formed by the prior belief and the de-

³⁶ Due to their probabilistic nature, probabilistic extreme values (zero or one) are theoretically not possible throughout the BDF, whereas empirical measures, given insufficient data, can result in such.

cision cost policy (effectively, operating points) which are compared to *objective* evidence estimates (LLR scores). The formal decoupling of *subjective* and *objective* domains accommodates for fully automated decision making systems (with formalized requirements, parameterizing thresholds) as well as for the optimization of *objective* evidence reporting independently of thresholds, e.g., by optimizing the *goodness of LLR scores* (C_{llr} as figure of merit) yielded by a recognition system. The first is suitable for commercial application scenarios, whereas the latter is suitable for forensic application scenarios: following the LLR estimation by the *forensic practitioner* (on the *source* and *activity levels*), the *trier of fact* assigns priors based on other elements of the case in order to make a decision (on the *offense level*).

The following section introduces the underlying concepts of the BDF, namely the total probability theorem, Bayesian inference, [strictly proper scoring rules](#), and effective operating points. Furthermore, BDF related performance visualizations are introduced, reporting on decision risk and the information supplied to decision making by a system.

2.4.1 Total Probability Theorem and Identity Inference

The *total probability theorem* states the calculus, which is necessary to outline and distinguish between propositions (prediction classes) [63, 64]. By denoting probabilities to the prior probability of each proposition, a *subjective degree of belief* is formalized in each proposition to be true. This prior belief is updated by a system's *objective* assessment of the evidence, resulting in the *subjective* posterior belief in the truth of each proposition. Based on the posterior probabilities, biometric identities are inferred, and informed decisions are made. The basis of the BDF is the *total probability theorem*: let \mathcal{U} partition the set \mathcal{U} by n *mutually exclusive*³⁷ and *exhaustive*³⁸ subsets A_1, \dots, A_n , such that their union equals \mathcal{U} [63, 64]:

$$\mathcal{U} = [A_1, \dots, A_n] . \quad (2.22)$$

In terms of biometric recognition, these partitions represent either verification outcomes or (closed-set) identification outcomes. Open-set identification outcomes can be targeted as well, when modeling the prior belief in unknown propositions (such as by uninformative priors). In this dissertation, LLRs inform on a binary decision (on \mathcal{A} or \mathcal{B}), such that *relevant populations* [104] need to be defined, employed for modeling the nominator $\Pr(E|\mathcal{A})$, and the denominator $\Pr(E|\mathcal{B})$ likelihood estimates that form an LLR $\log \frac{\Pr(E|\mathcal{A})}{\Pr(E|\mathcal{B})}$. All existing data that is relevant to the decision inference model is reflected by the set

³⁷ Mutually exclusive: $\{\} = \bigcap \{A_i | A_i \in \mathcal{U}\}$.

³⁸ Exhaustive: $\mathcal{U} = \bigcup \{A_i | A_i \in \mathcal{U}\}$.

\mathcal{U} ; the *relevant population* for class \mathcal{A} and the *relevant population* for class \mathcal{B} (and of other classes) result from a partition \mathcal{U} of \mathcal{U} .

The BDF requires the *relevant class populations* to be *mutually exclusive* and to be *exhaustive* in order to answer a *simple question* within a *decision inference chain* (e.g., forensic case work might link more evidence than biometric). Notably, the BDF is a model to decision making, and an aspect of the real world is modeled in consequence (not the entirety of the real world). This particular detail is crucial in two ways: one might argue that the *relevant population* was capable of helping to answer a stated *simple question*, which would be ideal if the depending argumentation remained feasible. Another, however, might argue that the *simple question* was answerable by the chosen *relevant population*, which would be practicable if the depending argumentation followed best practices. For the first argument, the stated *simple question* needs to be addressable, e.g., by the forensic practitioner employing a biometric system; for the latter, the chosen *relevant population* needs to be reflective of the current case, e.g., to be justifiable to a *trier of fact*. Therefore, the choice of how and which \mathcal{U} partitions \mathcal{U} is crucial.



Figure 2.6: Total probability theorem example with subset partitions $\mathcal{U}_i = [A_1, \dots, A_n]$ (black) and $\mathcal{U}_{ii} = [B_1, B_2, B_3]$ (green, blue, red).

Fig. 2.6 illustrates two partition examples $\mathcal{U}_i, \mathcal{U}_{ii}$ of \mathcal{U} :

- The partition \mathcal{U}_i depicts an arbitrary large number of n subsets with, e.g., *relevant populations* as A_1 for modeling class \mathcal{A} and A_2, \dots, A_n for modeling class \mathcal{B} as *not class \mathcal{A} in \mathcal{U}* , where the subsets are *mutually exclusive* and *exhaustive* in terms of the set \mathcal{U} . However, this partition is prone to run in intractable requirements: the stated *simple question* would require to reflect the entire world, which is deemed to be unaddressable (independently of whether or not one employs a sophisticated decision framework). This partition approach, however, might appeal to commercial application scenarios, when seeking further modeling generalizations.
- The partition \mathcal{U}_{ii} comprises 3 subsets as $B_1 = \{A_1, A_4\}$, $B_2 = \{A_2, A_3\}$, $B_3 = \{\{A_i \mid i \in \mathbb{N}, 4 < i < n\}, A_n\}$ with, e.g., *relevant populations* as B_1 for modeling class \mathcal{A} , B_2 for modeling class \mathcal{B} , and B_3 for modeling other populations, also being *mutually exclusive* and *exhaustive* subsets in terms of the set \mathcal{U} . However,

this partition limits the range of *simple questions* asked by the *trier of fact*, e.g., the subset A_1 and A_4 could reflect the voice of close relatives of the same gender³⁹ and the subsets A_2 and A_3 could reflect the voices of two other suspects than the one represented by A_1 . As a benefit, these *simple questions* are addressable by, e.g., a forensic practitioner. Notably, one needs to bear in mind that the weight of evidence will solely report on B_1 and B_2 , summarized by the LLR, and B_3 will remain without likelihood estimation (unless modeled by prior beliefs in the remaining classes and data availability to represent their relevant population). In other words, the *trier of fact* needs to examine the set \mathcal{U} , where forensic practitioners would report on the weight of evidence between the subsets B_1 and B_2 on the *source* and *activation levels*, leaving B_3 either to another evidence report within a case (e.g., the LLR of B_3 compared to the union of B_1 and B_2), or to further considerations carried out on the *offense level*.

Utilizing \mathcal{U} , the *total probability* $\Pr(E)$ of an evidence observation E (the new information) is partitioned by the subsets of \mathcal{U} [63, 64]:

$$\Pr(E) = \Pr(E|A_1) \Pr(A_1) + \dots + \Pr(E|A_n) \Pr(A_n), \quad (2.23)$$

where E is *arbitrary but fixed*.⁴⁰

Considering *binary decisions*, i.e., decisions of binary outcome, e.g., ($\{yes, no\}$, $\{true, false\}$, $\{accept, reject\}$), the proposition set $\mathcal{U} = [A, B]$ comprises two subsets (likewise two partitions) A, B , which can be part of a more complex $\mathcal{U} \subseteq \mathcal{U}$. For a *simple question* asked, however, the model of \mathcal{U} is sufficient, e.g., for A versus B ; A_1 versus A_2 ; or B_1 versus B_2 . Fig. 2.7a exemplarily depicts⁴¹ a two subset partitioning of $\Pr(E)$ in geometrical terms: the total probability $\Pr(E)$ is the combined area of the rectangles whose dimensions are the probability of each subset partitioned in \mathcal{U} (the prior beliefs: $\Pr(A), \Pr(B)$) and the

³⁹ The so-called *brother effect* is not only well-known to forensic practitioners. It describes that the voices of closely related male speakers sound similar on the telephone.

⁴⁰ In [63], the new information is referred to as *event* rather than *evidence*. In the terminology of *Bayesian inference*, the term *evidence* is preferred. In contrast to *specimens*, i.e., a sample of a class, *evidence* is data presented in proof of the facts and generalizes to any data derived from or as the specimen itself. Conventionally, the *total probability* of an evidence $\Pr(E)$ reflects its emission probability. In other words, the same (*fixed*) observed evidence E can be emitted by any subset $A_1, \dots, A_n \in \mathcal{U}$, but with varying emission probabilities $\Pr(E|A_1), \dots, \Pr(E|A_n)$, depending on the particular subset. The *law of total probability* considers the prior probability of each subset $\Pr(A_1), \dots, \Pr(A_n)$, such that the union of all pairwise disjoint subset-dependent evidence emission probabilities resembles the *total probability* of the evidence. This holds true for any measurable evidence observation (*arbitrary*).

⁴¹ The figure is inspired by educational videos of the YouTube *3Blue1Brown* channel by Grant Sanderson, see: https://www.youtube.com/channel/UCYO_jab_esuFRV4b17AJtAw.

depending probability of E assuming that solely one of the subsets is true (the conditional probabilities of the LR: $\Pr(E | \mathcal{A})$, $\Pr(E | \mathcal{B})$). As all subsets are mutually exclusive (each subset representing one among mutually exclusive propositions), solely one of the propositions of the subsets can be actually true.

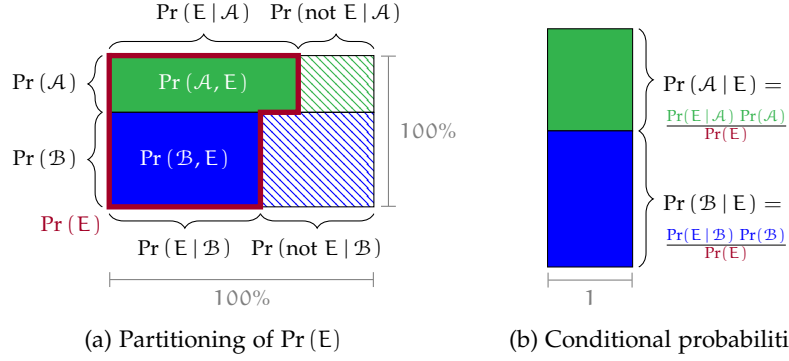


Figure 2.7: Partitioning of the total probability depending on beliefs, leading to conditional probabilities of the *Bayes rule*.

Fig. 2.7b shows how the posterior belief is derived as the conditional probability $\Pr(A_i | E)$ for a subset A_i given the evidence E [63, 64]:

$$\begin{aligned} \Pr(A_i | E) &= \frac{\Pr(A_i, E)}{\Pr(E)} \\ &= \Pr(E | A_i) \frac{\Pr(A_i)}{\Pr(E)}, \end{aligned} \quad (2.24)$$

with the joint probability $\Pr(A_i, E)$, i.e., of A_i and E being true:

$$\Pr(A_i, E) = \Pr(A_i | E) \Pr(E) = \Pr(E | A_i) \Pr(A_i). \quad (2.25)$$

The Bayes' theorem follows [63, 64]:

$$\Pr(A_i | E) = \frac{\Pr(E | A_i) \Pr(A_i)}{\Pr(E | A_1) \Pr(A_1) + \dots + \Pr(E | A_n) \Pr(A_n)}. \quad (2.26)$$

Example: Poker (Texas Hold'em), Straight versus Flush

Let's imagine you are playing poker, and there is one opponent left in the game.^a Flush and turn are drawn: two low cards (one of hearts), ten of spades, queen, and ace of hearts. You are holding jack and king of spades, a straight with the cards on the table. Only a flush can beat you, a flush of hearts.

What is the probability of your opponent holding two cards of hearts, making it five cards of hearts with the three on the table, a flush? There are 10 hearts of 45 cards remaining, your opponent is holding two cards. For your opponent holding a flush (proposition \mathcal{A}), the prior probability is $\Pr(\mathcal{A}) = \frac{\binom{10}{2}}{\binom{45}{2}} = \frac{45}{990} \approx 4.5\%$; for your

opponent not holding a flush (proposition \mathcal{B}), the prior probability is $\Pr(\mathcal{B}) = 1 - \Pr(\mathcal{A}) = \frac{945}{990} \approx 95.5\%$. *Something is happening!*

Your opponent places a bet—a high bet—a bluff? How does it change *your belief in her cards*? Let's examine the evidence E .

In the light of her having a flush (\mathcal{A} is true), a high bet is reasonable, let $\Pr(E|\mathcal{A}) = 97\%$ be the model output in this example. If she is not holding a flush (\mathcal{B} is true), she is bluffing by placing a high bet. For this example, let the model of her play style in the particular game situation return $\Pr(E|\mathcal{B}) = 30\%$. The joint probabilities are $\Pr(\mathcal{A}, E) = \frac{97}{2200}$, $\Pr(\mathcal{B}, E) = \frac{630}{2200}$, and the total probability of the evidence is $\Pr(E) = \Pr(\mathcal{A}, E) + \Pr(\mathcal{B}, E) = \frac{727}{2200} \approx 33\%$. Your prior beliefs are updated by the new information to the posterior beliefs: $\Pr(E|\mathcal{A}) = \frac{97}{727} \approx 13.3\%$, $\Pr(E|\mathcal{B}) = 1 - \Pr(E|\mathcal{A}) = \frac{630}{727} \approx 86.7\%$. The high bet increased the plausibility of her holding a flush from 4.5% to 13.3% by a factor of about 3. Let's now return to speaker recognition in unconstrained environments.

^a This example is adapted from Grant Sanderson's video preview *Bayes' rule!* on [patreon.com](https://www.patreon.com), 2017-06-15.

As the emphasis of this dissertation is placed on binary decisions, i.e., biometric verification, \mathcal{U} is effectively partitioned into two subsets representing the competing propositions of a biometric claim:

\mathcal{A} : *same subject (true biometric identity claim)*; versus

\mathcal{B} : *not $\mathcal{A} \Leftrightarrow$ different subjects (false biometric identity claim)*.

In the BDF [56, 57], conditional probabilities are employed, such that by taking the ratio of the Bayes rules for either posterior probabilities, the $\Pr(E)$ term cancels out, and the posterior ratio $\frac{\Pr(\mathcal{A}|E)}{\Pr(\mathcal{B}|E)}$ depicts the decision boundary solely regarding the LR $\frac{\Pr(E|\mathcal{A})}{\Pr(E|\mathcal{B})}$ and the prior ratio $\frac{\Pr(\mathcal{A})}{\Pr(\mathcal{B})}$:

$$\frac{\Pr(\mathcal{A}|E)}{\Pr(\mathcal{B}|E)} = \frac{\Pr(E|\mathcal{A}) \frac{\Pr(\mathcal{A})}{\Pr(E)}}{\Pr(E|\mathcal{B}) \frac{\Pr(\mathcal{B})}{\Pr(E)}} = \frac{\Pr(E|\mathcal{A})}{\Pr(E|\mathcal{B})} \frac{\Pr(\mathcal{A})}{\Pr(\mathcal{B})}. \quad (2.27)$$

For the purpose of simplification, the value⁴² of the prior probability ratio $\frac{\Pr(\mathcal{A})}{\Pr(\mathcal{B})}$ is summarized by a scalar π that reflects the effective prior probability of class \mathcal{A} and is thus referred to as the *target prior* (targeting class \mathcal{A} over class \mathcal{B}). Eq. (2.27) is simplified regarding the *target prior* π and the *target posterior* $\Pr(\mathcal{A}|E)$ [28, 79]:

$$\frac{\Pr(\mathcal{A})}{\Pr(\mathcal{B})} = \frac{\pi}{1 - \pi} \quad \Leftrightarrow \quad \pi = \frac{\Pr(\mathcal{A})}{\Pr(\mathcal{A}) + \Pr(\mathcal{B})},$$

⁴² Prior probability values of a class (e.g., class \mathcal{A}) can be assigned in terms of, e.g., (a) the expected occurrence proportion of that class within the dataset, (b) *maximum entropy* (with the value 0.5 for two classes), and (c) prior distributions that can be *informative* (e.g., conjugate prior distributions) or *uninformative* (e.g., Jeffrey's prior).

$$\frac{\Pr(\mathcal{A} | \mathcal{E})}{\Pr(\mathcal{B} | \mathcal{E})} = \frac{\Pr(\mathcal{A} | \mathcal{E})}{1 - \Pr(\mathcal{A} | \mathcal{E})} = \frac{\Pr(\mathcal{E} | \mathcal{A})}{\Pr(\mathcal{E} | \mathcal{B})} \frac{\pi}{1 - \pi}. \quad (2.28)$$

By summarizing the prior probability impact in terms of π , one can easily infer *exhaustiveness* within the model (from $\frac{\pi}{1-\pi}$): the prior belief is updated regarding the targeted class (class \mathcal{A}) and its opposing class (class \mathcal{B}). The decision model answers a *simple question*, i.e., between the two classes \mathcal{A} and \mathcal{B} . As a parameterization of π for any parameterizations of $\Pr(\mathcal{A})$ and $\Pr(\mathcal{B})$ will always exist, the BDF operates in an *exhaustive* set.

Notably, the BDF infers identities by the *subjective* posterior probabilities, where a decision maker formalizes requirements on the posterior probability ratio $\frac{\Pr(\mathcal{A} | \mathcal{E})}{1 - \Pr(\mathcal{A} | \mathcal{E})}$. By separating the *subjective* prior and posterior from the *objective* LR, Eq.(2.28) is reformulated as:

$$\frac{\Pr(\mathcal{A} | \mathcal{E})}{1 - \Pr(\mathcal{A} | \mathcal{E})} \frac{1 - \pi}{\pi} = \frac{\Pr(\mathcal{E} | \mathcal{A})}{\Pr(\mathcal{E} | \mathcal{B})}, \quad (2.29)$$

such that the decision layer (left side) is decoupled from the system output (right side). The following section elaborates on the formal definition of the decoupled decision layer.

2.4.2 Decoupled Decision Layer

Decisions are made by examining the posterior ratio $\frac{\Pr(\mathcal{A} | \mathcal{E})}{1 - \Pr(\mathcal{A} | \mathcal{E})}$ regarding a decision policy. Propositions and decision outcome depending costs are employed, conceptually regarding a required impact of $\Pr(\mathcal{A} | \mathcal{E})$. A cost matrix associates two costs to each proposition: a *success cost* and an *error cost*. When targeting binary decisions and assigning costs zero to successes, solely Type I and Type II error costs c_I, c_{II} need to be considered:

$$\begin{aligned} c_I &: \text{cost for falsely accepting proposition } \mathcal{B}; \\ c_{II} &: \text{cost for falsely rejecting proposition } \mathcal{A}. \end{aligned}$$

Costs are incentives and define the required impact of posterior odds to decision making. In this dissertation, costs are treated in a formalized but applicable manner, such that different perspectives on costs are denoted as⁴³:

- $c_{\mathcal{A}}, c_{\mathcal{B}}$: the costs of erroneous decisions against the propositions \mathcal{A}, \mathcal{B} , which can be associated with common factors and comprise units (e.g., expressed in € or \$). Eventually, solely the cost ratio is of interest, units and common factors cancel out;

⁴³ In the speaker recognition community, priors and costs are addressed in terms of $\pi, c_I, c_{II}, \tilde{\pi}$. The definitions of $c_{\mathcal{A}}, c_{\mathcal{B}}, c$ are employed here for the purpose of easier tractability for readers with another background, such as image based biometrics, machine learning or banking. The outline and definition of the parameter c might be seen as a contribution of this dissertation (bridging gaps).

- c_I, c_{II} : Type I and Type II error costs (without units);
- c : a scalar summary of the cost ratio (its probabilistic form);

where these three concepts relate as:

$$c = \frac{c_{II}}{c_{II} + c_I} = \frac{c_{\mathcal{A}}}{c_{\mathcal{A}} + c_{\mathcal{B}}} . \quad (2.30)$$

Example: Decision Costs in Smart Home Applications I/III

Voice activated smart home applications, e.g., lighting control or home security access, range from non-critical to critical/sensitive:

- Addressing the costs of erroneous decisions in lighting control, one might aim at energy consumption versus inconvenience costs and define $c_{\mathcal{B}} = 20$ cents for a Type I error and $c_{\mathcal{A}} = 40$ cents for a Type II error. As such, $c_I = 1$, $c_{II} = 2$, and $c = \frac{2}{3}$.
- Regarding the costs of erroneous decisions in home security access, one might target liability versus inconvenience costs and define $c_{\mathcal{B}} = 99\,000$ € for a Type I error and $c_{\mathcal{A}} = 100$ € for a Type II error. As such, $c_I = 990$, $c_{II} = 1$, and $c = \frac{1}{991}$.

The definition of costs is a subjective belief, so is *any* decision policy. Different **biometric system operators** might require different cost constraints from a system **biometric system vendor** for similar application scenarios.

Given an evidence E , binary Bayes decisions are made by minimizing the Bayes risk, i.e., regarding the cost notations and posteriors [28]:

$$\begin{aligned} \Pr(\mathcal{B}|E) c_{\mathcal{B}} &\leq \Pr(\mathcal{A}|E) c_{\mathcal{A}}, & \Pr(\mathcal{B}|E) c_I &\leq \Pr(\mathcal{A}|E) c_{II}, \\ (1 - \Pr(\mathcal{A}|E)) (1 - c) &\leq \Pr(\mathcal{A}|E) c, \\ \frac{c_{\mathcal{B}}}{c_{\mathcal{A}}} &\leq \frac{\Pr(\mathcal{A}|E)}{1 - \Pr(\mathcal{A}|E)}, & \frac{c_I}{c_{II}} &\leq \frac{\Pr(\mathcal{A}|E)}{1 - \Pr(\mathcal{A}|E)}, & \frac{1 - c}{c} &\leq \frac{\Pr(\mathcal{A}|E)}{1 - \Pr(\mathcal{A}|E)}. \end{aligned} \quad (2.31)$$

Combining Eqs. (2.29) and (2.31), comparison (the LR $\frac{\Pr(E|\mathcal{A})}{\Pr(E|\mathcal{B})}$) and decision layers (the LR threshold τ) are decoupled as the posterior ratio is substituted with the cost ratio:

$$\tau = \frac{1 - c}{c} \frac{1 - \pi}{\pi} \leq \frac{\Pr(E|\mathcal{A})}{\Pr(E|\mathcal{B})}. \quad (2.32)$$

For simplification and mathematical convenience, the prior, posterior, and cost ratios are expressed in terms of their log-odds form, an alternative way of expressing probabilities. Thereby, logarithmic

Bayes thresholds η are denoted as **log-likelihood ratio (LLR)** operating points (LLR thresholds), such that the log-odds of the *target prior* are re-biased by the log-odds of the *decision costs* (and vice versa):

$$\begin{aligned}\eta = \log \tau &= \log \frac{1-c}{c} \frac{1-\pi}{\pi} \\ &= \text{logit}(1-c) + \text{logit}(1-\pi) \\ &\leq \log \frac{\Pr(E|\mathcal{A})}{\Pr(E|\mathcal{B})},\end{aligned}\tag{2.33}$$

where the *logit* transformation (to log-odds) is defined as⁴⁴:

$$\text{logit}(x) \equiv \log \frac{x}{1-x},\tag{2.34}$$

being its inverse function, the *sigmoid* function:

$$\sigma(x) \equiv (1 + e^{-x})^{-1}.\tag{2.35}$$

Notably, the *sigmoid* function is a particular case of the *logistic* function $F(x)$, specified by location and scale parameters μ, s :

$$F(x) = \left(1 + e^{-\frac{x-\mu}{s}}\right)^{-1}.\tag{2.36}$$

The *logit* function represents the log-odds associated with probability $x \in (0, 1)$. Notably, log-odds are fundamental to *logistic regression* (to the *logit* model). By employing log-odds, the identity inference and decision making processes are simplified.

Example: Decision Costs in Smart Home Applications II/III

Voice activated smart home applications, e.g., lighting control or home security access, operate in an environment with few subversive interaction attempts, see section 2.1.1. Here, the exemplary occurrence of a true biometric identity claim is ten times more likely than a false biometric identity claim. The *target prior* is $\pi = \frac{10}{11}$, i.e., $\frac{\pi}{1-\pi} = 10$ in odds notation. The depending log-odds are $\text{logit}(\pi) = \text{logit}\left(\frac{10}{11}\right) = \log \frac{\pi}{1-\pi} = \log(10)$.

Depending Bayes thresholds η are derived as the log of the ratio between cost and prior odds, which in terms of log-odds is formalized by the (symmetric) sum: $\eta = \text{logit}(1-c) + \text{logit}(1-\pi)$, see Eq. (2.33). In these smart home application examples, the threshold is partially defined by the prior log-odds: $\eta = \text{logit}(1-c) - \log(10)$, where the cost log-odds ($\text{logit}(1-c)$) depend on the decision costs of an application:

- For lighting control with $c = \frac{2}{3}$, the cost log-odds are $\text{logit}(1-c) = -\log(2)$, such that the Bayes threshold is $\eta = -\log(2) - \log(10) = -\log(20) \approx -3.0$, i.e., -3.0 is the smallest value of LLR scores resulting in the favor of proposition \mathcal{A} .

⁴⁴ Notably, $-\text{logit}(x) = \log \frac{1-x}{x} = \text{logit}(1-x)$ and $\eta = \text{logit}(1-c) + \text{logit}(1-\pi)$.

- Regarding home security access with $c = \frac{1}{991}$, the cost log-odds are $\text{logit}(1 - c) = \log(990)$, such that the Bayes threshold is $\eta = \log(990) - \log(10) = \log(99) \approx 4.6$.

Notably, zero valued LLRs represent equality in the support of either propositions \mathcal{A}, \mathcal{B} (e.g., $0 = \log 1 = \log \frac{0.5}{0.5} = \log \frac{50\%}{50\%}$). Thus, depending zero valued LLR thresholds require at least as much support of proposition \mathcal{A} as of proposition \mathcal{B} but reflect operating points of *maximum entropy* regarding costs and priors.^a Regarding the lighting control scenario, LLRs greater than or equal to $\eta = -3.0$ are required. This accommodates not only for LLRs supporting security decisions (LLRs greater than zero), but also includes LLR thresholds $\eta \in [-3.0, 0.0]$, relating to LLRs supporting proposition \mathcal{B} over \mathcal{A} . In other words, the depicted lighting control application naturally accommodates for LLRs favoring convenience and security decisions but not for extremely high convenience decisions (e.g., represented by LLR thresholds $\eta \ll -5.0$). Vice versa, in the home security access scenario, LLR thresholds $\eta \geq 4.6$ are required, excluding LLRs favoring lower security decisions (corresponding to LLR thresholds $0 < \eta < 4.6$). The depicted home security access application naturally accommodates for LLRs favoring higher security decisions only.

^a The cost requirement and the prior probability at *maximum entropy* are 0.5, and the depending impact on the LLR threshold resembles $\text{logit}(0.5) = 0$ log-odds.

As seen in Eq. (2.33), the ratio $\frac{C_{\mathcal{B}}}{C_{\mathcal{A}}} \rightarrow \infty$ if $c \rightarrow 0$, and $\frac{C_{\mathcal{B}}}{C_{\mathcal{A}}} \rightarrow 0$ if $c \rightarrow 1$ is analogous to the (reciprocal) prior ratio $\frac{\Pr(\mathcal{B})}{\Pr(\mathcal{A})}$ and π . Thus, under the typical assumption of positive costs, the pair of (c, π) will define the whole range of possible *applications* of the system. A particular application represents a possible scenario, where a system using LLRs will be operating, determined by a particular pair (c, π) and thus a particular value of η by Eq. (2.33). Seeking further scalar summaries of the operative decision policies (c, π) , an *effective prior* $\tilde{\pi}$ (likewise, an *effective cost* or *effective policy*) is denoted as [28, 79]:

$$\begin{aligned}\tilde{\pi} &= \frac{\pi c}{\pi c + (1 - \pi)(1 - c)}, \\ \eta &= \text{logit}(1 - \tilde{\pi}) \quad \text{with decision making by:} \\ \eta &\leq \log \frac{\Pr(E|\mathcal{A})}{\Pr(E|\mathcal{B})}.\end{aligned}\tag{2.37}$$

Example: Decision Costs in Smart Home Applications III/III

Effective priors bridge the gap between *target priors* π and *decision costs* c , as formal application values (c, π) are mappable to a specific $\tilde{\pi}$ value and thus between one another. For example, costs remain unspecified in forensic scenarios, assuming unit costs, i.e., maxi-

imum cost entropy: $c = \frac{1}{2}$. By utilizing *effective priors*, commercial and forensic applications (c, π) are interrelated:

- For the lighting control application $(\frac{2}{3}, \frac{10}{11})$ with $\eta = -\log(20)$, the effective prior resembles $\tilde{\pi} = \frac{20}{21}$, also resembling applications like $(\frac{1}{2}, \frac{20}{21})$, $(\frac{20}{21}, \frac{1}{2})$, $(\frac{1}{100}, \frac{1980}{1981})$, $(\frac{1980}{1981}, \frac{1}{100})$.
- For the home security access application $(\frac{1}{99}, \frac{10}{11})$ with $\eta = \log(99)$, the effective prior resembles $\tilde{\pi} = \frac{1}{100}$, also resembling applications like $(\frac{1}{2}, \frac{1}{100})$, $(\frac{1}{100}, \frac{1}{2})$, $(\frac{20}{21}, \frac{1}{1980})$, $(\frac{1}{1980}, \frac{20}{21})$.

Notably, the effective prior $\tilde{\pi} = 0.01$ has been employed as one application configuration in the biannual [speaker recognition evaluation \(SRE\)](#) series of the [US National Institute of Standards and Technology \(NIST\)](#) since 2010, the major research challenge of the speaker recognition community.

For the purpose of training recognition systems for multiple applications, i.e., application-independently, logistic regression models, among others, optimize the effective prior $\tilde{\pi}$ as a scalar representation instead of more complex operating point representations in order to yield well-calibrated system outputs ([LLR](#) scores) over the range of all possible applications $\{(c, \pi) \mid c \in (0, 1), \pi \in (0, 1)\}$.⁴⁵

2.4.3 Decision Risk Performance

The performance of systems in specific applications is examined by the Bayes risk [59]. Given a set of binary decision scores S , the *empirical Bayes risk* \mathcal{R} is the expected cost value, a weighted sum of the Type I and Type II error rates (e.g., [false match rate \(FMR\)](#), [false non-match rate \(FNMR\)](#)). Weights resemble the depending decisions costs (e.g., c_A, c_B ; c_I, c_{II} ; or c) and the depending prior probabilities (e.g., $\Pr(\mathcal{A})$, $\Pr(\mathcal{B})$; or π). One might define the *Bayes risk* as either $\mathcal{R}(S \mid \Pr(\mathcal{A}), \Pr(\mathcal{B}), c_A, c_B)$ or in terms of the target prior and unit-free costs as $\mathcal{R}(S \mid \pi, c_I, c_{II})$ or in terms of the summarized priors and costs as $\mathcal{R}(S \mid \pi, c)$:

$$\begin{aligned} \mathcal{R}(S \mid \Pr(\mathcal{A}), \Pr(\mathcal{B}), c_A, c_B) &= \Pr(\mathcal{A})c_A \text{FNMR}(\eta) + \Pr(\mathcal{B})c_B \text{FMR}(\eta) , \\ \mathcal{R}(S \mid \pi, c_I, c_{II}) &= \pi c_{II} \text{FNMR}(\eta) + (1 - \pi) c_I \text{FMR}(\eta) , \\ \mathcal{R}(S \mid \pi, c) &= \pi c \text{FNMR}(\eta) + (1 - \pi) (1 - c) \text{FMR}(\eta) , \end{aligned} \quad (2.38)$$

assigning zero costs to successful decisions. Notably, the definition of an application's operating point, i.e., $(c_A, c_B, \Pr(\mathcal{A}), \Pr(\mathcal{B}))$; (c_I, c_{II}, π) ; (c, π) ; or $(\tilde{\pi})$ also parameterizes the Bayes threshold η . In the [Bayesian](#)

⁴⁵ For $c = 0, c = 1, \pi = 0, \pi = 1$, LLR scores remain without impact, i.e., decisions are already made without any employment of a recognition system, as always one of the propositions \mathcal{A} or \mathcal{B} is favored per default in decision policies with $\eta = \pm\infty$.

decision framework (BDF), the distribution of LLR scores ideally corresponds to the class proportion at the depending Bayes threshold values over the entire range of possible applications. As such, the Bayes risk is the expected (average) posterior decision cost: Type I and Type II error rates resemble the expected proportion of erroneous decisions, occurring with a prior probability and causing a cost impact of making an erroneous decision. The sum of the class-dependent expected risks resembles the expected risk value over all classes (due to the linearity of the *expectation operator*). For an arbitrary but fixed parametrization, the Bayes risk is referred to as **decision cost function (DCF)**.⁴⁶

The parameterization of the empirical Bayes risk is simplified by employing the effective prior $\tilde{\pi}$, such that the evaluation criterion is (merely) scaled to the *empirical Bayes error rate* \mathcal{E} (as a DCF) [28, 79]:

$$\begin{aligned}\mathcal{E}(S|\tilde{\pi}) &= \mathcal{R}(S|\tilde{\pi}, 1, 1) = \frac{\mathcal{R}(S|\pi, c_I, c_{II})}{\pi c_{II} + (1 - \pi) c_I} \\ &= \tilde{\pi} \text{FNMR}(\eta) + (1 - \tilde{\pi}) \text{FMR}(\eta) .\end{aligned}\quad (2.39)$$

Figs. 2.8a to 2.8b visualize empirical Bayes error rates for the exemplary systems in Figs. 2.3a to 2.3c among effective priors $\tilde{\pi}$, which are referred to as **applied probability of error (APE)** plots⁴⁷, cf. [28, 61]. APE plots assess DCFs over all possible application parameterizations, depicting the *actual decision risk* compared to the *default decision risk* (if all LLR scores are equal to zero) and to the *discrimination power* a system is potentially capable of supporting if scores of the system were ideally calibrated, i.e., being LLRs.

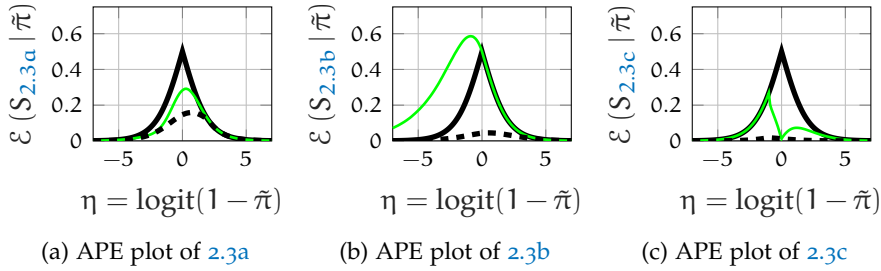


Figure 2.8: Examples of APE plots for synthetic score distributions: for score sets of Figs. 2.3a–2.3c with class \mathcal{B} scores (red), class \mathcal{A} scores (green); APE plots (a), (b), (c) with $\mathcal{E}(S|\tilde{\pi})$ (green), $\mathcal{E}_{\min}(S|\tilde{\pi})$ (dashed), and the default performance $\mathcal{E}(S_0|\tilde{\pi})$ (black).

However, even when based on Bayesian statistics, systems might yield badly calibrated scores, e.g., in the light of insufficient data

⁴⁶ DCFs are employed as the primary evaluation criterion within the NIST SREs.

⁴⁷ APE plots depict performance depending on the effective prior, i.e., the negative Bayes threshold $-\eta = \text{logit} \tilde{\pi}$. For harmonization purposes, one may choose to present the x-axis regarding η , such that (LLR) score histograms are co-aligned.

amounts or poor data quality, such that one would be interested to recalibrate these scores to LLRs. In single application scenarios, the calibration to accommodate all possible LLR thresholds is not required as long as the lowest decision risk results at the depending Bayes threshold, as this is the only LLR value where an idealistic *actual decision risk* is necessary (for a single application).

To reveal the *discrimination power*, the DCF metric is exploited. By interpreting the DCF performance as a threshold-independent function and by sweeping over all threshold values while the effective prior weight remains fixed, the optimal DCF performance is revealed for each operating point $\tilde{\pi}$. The optimal DCF performance, **minimum DCF (minDCF)**, is referred to as \mathcal{E}_{\min} , representing the performance of a well-calibrated system⁴⁸ [79]:

$$\text{minDCF}(S|\tilde{\pi}) = \min_{-\infty \leq \eta \leq \infty} \tilde{\pi} \text{FNMR}(\eta) + (1 - \tilde{\pi}) \text{FMR}(\eta) . \quad (2.40)$$

In the following, *actual* performance will refer to empirical measurements with *minimum* referring to the performance on the assumption of well-calibrated system outputs (LLRs). The *minimum* resembles the *discrimination power*, whereas the *actual* performance reflects the *operational power*, i.e., the combination of the *discrimination power* and the *calibration loss*.

For the purpose of better visual comparability in which a similar visual distance conveys the same relative DCF impact in the tails as in the center of APE plots (sacrificing equal absolute DCF differences), [79] proposes a normalization of the empirical Bayes error rate utilizing a *default system*, emitting the score set S_0 . An S_0 default system yields the reference performance:

$$\mathcal{E}(S_0|\tilde{\pi}) = \min(\tilde{\pi}, 1 - \tilde{\pi}) , \quad (2.41)$$

where all LLRs equal zero; a system contributing as much information to decision making as a *coin toss*. The visualization gap of APE plots manifests in the DCF of the default system: distances to the default performance visually collapse in the tails of the APE plot, i.e., for applications of high convenience or of high security (e.g., $\eta = -5$ and $\eta = +5$). Fig. 2.9 illustrates the **normalized Bayes error rate (NBER)** plots, normalizing Fig. 2.8. Thereby, the contribution of FMR and FNMR error rates to the *minimum* and *actual* performance are depicted depending on the Bayes threshold η , cf. [79]. Conventionally, APE and NBER plots refer to $\text{logit}(\tilde{\pi}) = -\eta$ on x-axes, placing emphasis on the effective target prior due to forensic perspectives. In this dissertation, however, η is directly depicted on the x-axes, which might be more intuitive to layman examiners, seeking coherence within the performance assessment of similarity scores.

⁴⁸ In this sense, systems are numerically calibrated by exploiting all possible configurations under the preservation of monotonicity.

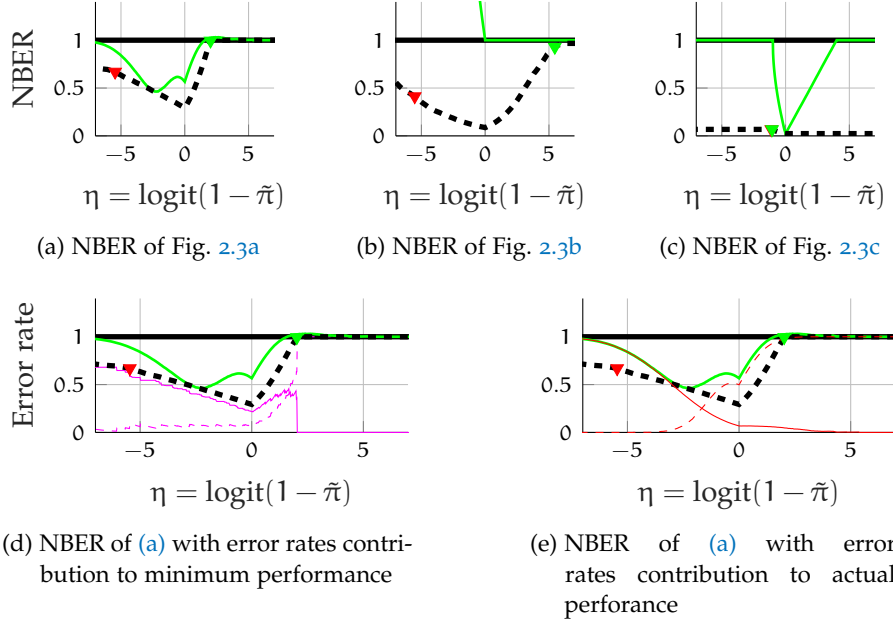


Figure 2.9: Examples of NBER plots with actual (green), minimum (dashed), and default (black) performance; [rule of 30](#) bounds are depicted dependent on effective priors $\tilde{\pi}$ regarding 30 false matches (green triangle) and 30 false non-matches (red triangle); the contribution depending on error rates is exemplarily depicted in (d), (e) in terms of $\tilde{\pi}$ -weighted FMR (dashed), FNMR (solid) contributions to the minimum (magenta) and to the actual DCF (red).

NBER plots application-dependently interrelate DCFs and error rates across different applications (c, π) , which are summarized in terms of the effective prior $\tilde{\pi}$. The exemplary NBER plots serve to illustrate that Bayes risk values depend on the formalized application's operating point: targeting cross-application use cases, performance assessment solely based on error rates will mislead decision making and increase operational costs, depending on the impact of erroneous decisions made. Thereby, score calibration is crucial so as to yield an operational performance close to the discrimination performance.

The goal of this dissertation is to assure low discrimination but also low calibration losses, targeting an application-independent system output, i.e., scores that can serve as LLRs for all operating points (e.g., for effective decision policies $\tilde{\pi}$ outlining Bayes thresholds η). In the following sections, the conventional error rate based performance assessment as well as score calibration are addressed. In forensic scenarios, score calibration and depending performance measures are motivated from *information theory* since priors and costs remain unknown or at maximum uncertainty during evidence reporting. Notably, these different perspectives on recognition performance are summarized by the same performance measure which emerges from either perspective and will serve as the primary evaluation metric here, namely C_{llr} .

and C_{llr}^{\min} . Interrelations between the different perspectives exist, particularly between the *Bayes risk*, *score calibration*, and *information gains*. Thereby, the *Bayes risk* is based on error rates (on conventional error rate performance) but outlines ideal *score calibration*: the decision risk is minimal when the score values are LLRs. Consequently, the conventional error rate trade-off assessment is interrelated with ideal score calibration. The following section depicts the *Bayes risk* from the perspective of conventional error rate trade-off assessment.

2.4.4 Error Rate Performance Visualization and Ideal Score Calibration

This section addresses the visual and numerical interrelation between: i) [detection error trade-offs \(DETs\)](#), ii) minDCFs, and iii) well-calibrated scores (LLRs). In terms of errors, binary decision performance is typically visualized by depicting Type II against Type I error rates, such as in DET plots [88]. DET plots are [receiver operating characteristic \(ROC\)](#) plots [90], ROCs plot class \mathcal{A} success rates against class \mathcal{B} error rates⁴⁹, with inverted y-axis, i.e., depicting the Type II error rate rather than the true positive rate. Furthermore, DETs warp both axes by the *probit* function, i.e., they interrelate probabilistic odds. By depicting empirical error rates in DETs, i.e., discrete and not continuous data, a steppy curve is obtained, representing the error trade-offs across all thresholds. As a complement to the steppy DET, calibrated scores can also be depicted in terms of a continuous interpolation over the underlying ROC. This interpolation is the [ROC's convex hull \(ROCCH\)](#). It visualizes minDCF values in the canvas of the y-axis inverted ROC. ROCCH and minDCF are closely related due to the [pool adjacent violators \(PAV\)](#)-LLR algorithm [28, 79, 105, 106]. Figuratively speaking, the ROCCH can be thought of as a rubber band put around the steppy ROC which yields a convex interpolation of continuous error trade-offs from the observed, empirical, discrete trade-offs in the ROC space. The ROCCH visualizes ideal score calibration (the PAV-LLR algorithm's output in the error rate trade-off domain) and relates to the minDCF as follows. As the minDCF continuously minimizes over thresholds, the ROCCH is a two-dimensional discrete minimization between ROC points in the $x, y \in \mathfrak{R}^2$ space (in the y-axis inverted [ROC](#) space) [79]:

$$[x, y] = \sum_{i=0}^n \alpha_i [\text{FMR}(i), \text{FNMR}(i)] \quad \text{with} \quad \alpha_i \geq 0, \sum_{i=0}^n \alpha_i = 1, \quad (2.42)$$

such that the minDCF (minimizing weighted error rates over thresholds) is interpreted as a discrete minimization over the vertices of the ROCCH. Notably, all minDCF $\tilde{\pi}$ weighting parameterizations resemble tangents to the ROCCH with a $\tilde{\pi}$ -depending slope where the

⁴⁹ ROCs were developed in the field of radar detection in the 1940s, gaining popularity in the 1960s.

value of a minDCF parameterization resembles the distance of a tangent to the plot's origin (geometric proofs are outlined in [66, 107]). An equivalence proof between the ROCCH algorithm and the PAV algorithm is shown in [108].

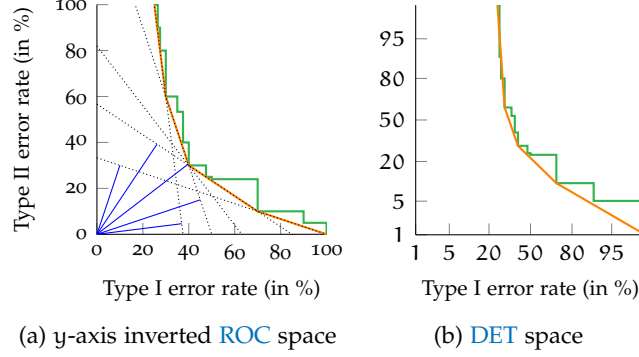


Figure 2.10: Example regarding DETs: steppy (green), ROCCH (orange), with (a) corresponding y-axis inverted ROC space: minDCF tangents (dotted) and lines (blue) with line lengths corresponding to minDCF values and (b) ROC and ROCCH in the DET space.

Fig. 2.10 illustrates a DET plot [88] with an exemplary steppy curve and its depending ROCCH curve. Notably, by swiping over all application parameterizations, i.e., over all Bayes thresholds, the maximum of all resembling minDCF values (across all $\tilde{\pi}$ weighting parameterizations) corresponds to the **equal error rate (EER)** of the ROCCH [28, 61, 79], which is referred to here as the EER. Fig. 2.11 exemplarily depicts the relationship between minDCF and the EER. Alternatively to the ROCCH, *isotonic regression* can be used to minimize the mean square error between raw scores and the fit to probabilistic class labels between 0, 1, i.e., the PAV-LLR algorithm [28]. For the sake of easy tractability, the PAV-LLR algorithm is referred to as PAV.

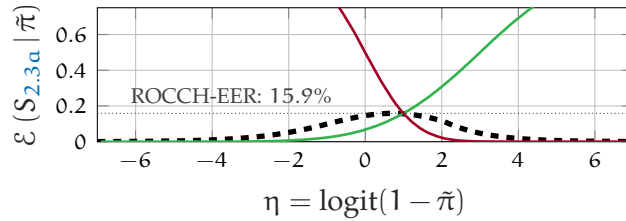


Figure 2.11: EER as maximum minDCF, cf. Fig. 2.8a, with minDCF (dashed), FMR (red), FNMR (green).

The PAV algorithm is elaborately examined⁵⁰ in [28] with motivation for speaker and language recognition systems. It can be easily expanded further to any binary or multi-class decision system. As depicted in [79], PAV calibrated scores resemble the ROCCH when

⁵⁰ Implementations to system evaluation, calibration, and fusion are found in [79].

visualizing Type II against Type I error rates. In other words, when depicting the discrimination performance of binary decision scores, the post-evaluation calibration of uncalibrated scores to LLRs is inherent to the nature of binary decision systems⁵¹, such that the performance visualization employing the ROCCH is the natural choice. The ROCCH is the *expected* ROC segment of (all) *optimistic* and *pessimistic* interpolations [90], where the [Bayesian decision framework \(BDF\)](#) furthermore employs the ROCCH in order to depict ideal decision making, i.e., after PAV calibration.

The PAV algorithm [28, 78] performs *isotonic regression* for two-class problems: PAV maps the entire range of score values to a unified score scale of posterior probabilities that are optimally calibrated. Fig. 2.12 shows the output of the PAV algorithm: a PAV value for each score is in the $[0, 1]$ domain, namely S_{PAV} , an optimally-calibrated posterior probability. The value is a posterior probability because PAV intrinsically considers the so-called *empirical prior-ratio* $\check{\pi}$ to compute its transformation, which is the proportion of class \mathcal{A} scores w.r.t. the total in a data set (a given⁵² value of π). A more detailed example is provided in annex B. Thus, S_{PAV} values should directly correspond to the log-odds of costs to make Bayes decisions, but decisions could only be made for $\check{\pi}$. Removing the log-odds influence of $\check{\pi}$, scores S are indeed LLR values [28, 79]:

$$S = \text{logit}(S_{PAV}) - \text{logit}(\check{\pi}) . \quad (2.43)$$

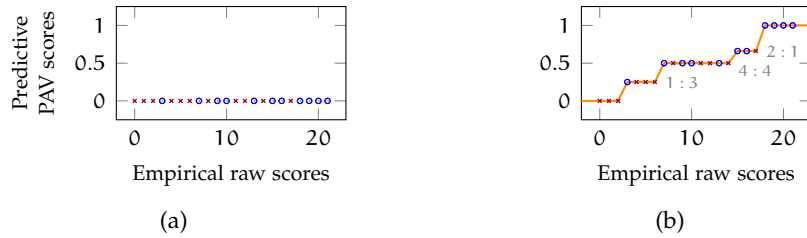


Figure 2.12: Example of PAV isotonic regression with scores of propositions \mathcal{A} (blue circles), \mathcal{B} (red crosses), and the resulting PAV function (orange). (a) Algorithm initialization (all scores to calibrate with 0 value of y-axis); (b) the predictive PAV function, preserving monotonicity: the y-axis increases with an increasing proportion of scores from class \mathcal{A} ; proportions as odds (gray).

As Eq. (2.40) provides a numerical solution to *score calibration* (the score-to-LLR transform), solely optimizing for one operating point, PAV rather poses an analytical solution to *score calibration*, providing

⁵¹ On the assumption of the total probability theorem, i.e., that the propositions to which the binary decision system is examining evidence are mutually exclusive but not necessarily exhaustive.

⁵² Notably, for evaluation purposes, an evaluator may choose a π suitable to the targeted application [28], not necessarily equaling $\check{\pi}$.

ideal LLRs for *all* operating points. Fig. 2.13 exemplarily illustrates the ideal LLRs corresponding to the synthetic scores of Figs. 2.3a to 2.3c. Notably, the PAV transform preserves error rate trade-offs. However, the depending histograms dramatically change as scores are transformed into their related ideal LLR representation. As such, the relation of thresholds and error rates changes, also changes. The experimental setup denotes bounds to empirical LLRs and thus to the possible shapes of error rate proportions. In this sense, proportions of (empirical) error rates are nothing but histograms of score distributions.

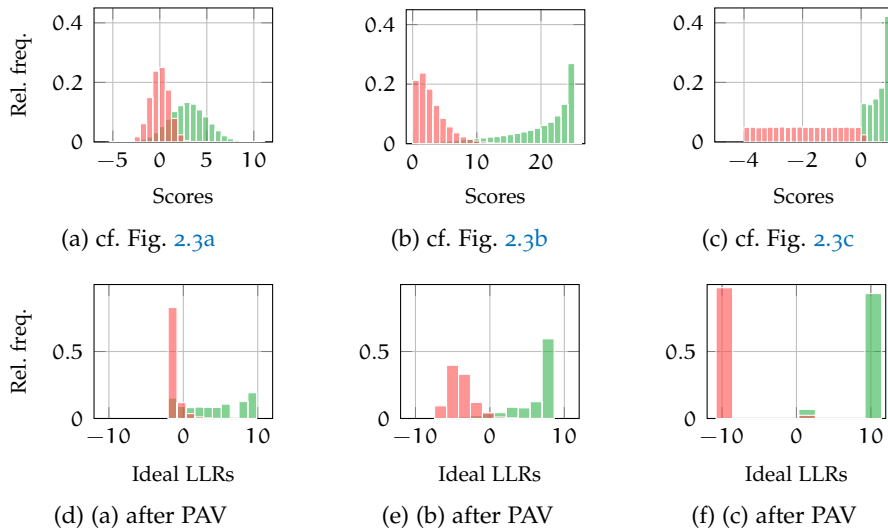


Figure 2.13: PAV-LLR example: (a) – (c) score histograms, (d) – (f) their ideal LLR representation: the latent subspace of *decision making*.

2.4.5 On Performance Visualizations in Forensic Evaluation

In forensic evaluative scenarios, the province of court⁵³ (decision making) is separated from the province of the forensic practitioner (reporting the weight of the evidence in terms of a score), such that the reports of a forensic witness are neither without bias towards prosecution nor towards defendant propositions, and the decision making is solely conducted by judges and jurors. As the forensic practitioner cannot know the operating point employed, the reported *weight of evidence* (the score) ideally provides accurate probabilistic predictions towards either proposition \mathcal{A} , \mathcal{B} for each operating point, which is evaluated by proper scoring rules regarding the *goodness* of scores to represent LLRs, cf. sections 2.3.1. In the 2015 methodological guidelines of the European network of forensic science institutes (ENFSI) [62],

⁵³ In investigatory scenarios, prosecutors or law enforcement make decisions. The point is that decisions are not made by the forensic practitioner.

the prediction accuracy is addressed in terms of the *probabilities of misleading evidence*. Effectively, these probabilities are FMR and $1 - \text{FNMR}$ values, which are ideally equal at the LLR threshold $\eta = 0$ (the depending system is well-calibrated at $\eta = 0$), where neither proposition is favored over the other. From the proportions of FMR and $1 - \text{FNMR}$ values, depending LLR values resemble across all thresholds, and the corresponding LLR thresholds express the required extend of which either proposition needs to be favored over another.

In forensics, the assessment perspective of recognition performance requires to shift from *the impact of decisions made by employing a biometric system* to *the benefit a decision maker has by employing a biometric system, regardless of any decision policy*. The former, while suitable in commercials, implies that the forensic practitioner makes decisions, which is in another province. The latter implies that the forensic practitioner reports the evidence (without any bias) in some way, ideally providing information to the province of court (the decision makers). For commercials, the forensic perspective can be interpreted as the analysis of recognition performance during research and system development. Therefore, the distribution of scores is examined in three different ways: (i) as the proportion of cases depending on LLR values—in commercials, error rates regarding PAV-LLR histograms (in the latent decision subspace)—; (ii) the separability of class instances (score distributions of [subjects](#)); and (iii) the expected separability resulting from all observed class instance—the *information gain* a system provides to decision making, particularly in comparison to employing no system reporting on the evidence.

Tippett plots [[109](#)] visualize the correspondence of LLR values to FMR and $1 - \text{FNMR}$ values. Limit Tippett plots [[110](#)] further depict lowest to highest error rates possible after score calibration (*universal bounds*), when considering the empirical prior $\hat{\pi}$ regarding either class \mathcal{A}, \mathcal{B} . Thereby, a pseudo PAV score calibration is conducted by assuming Gaussian distributions. Fig. [2.14](#) illustrates *Limit Tippett plots*: at $\eta = 0$, the vast gap between both *probabilities of misleading evidence* suggest badly calibrated scores. Compared to [DET](#) plots, Limit Tippett plots interrelate discrimination and calibration performance. Compared to [NBER](#) plots, the *universal bounds* depict the limitations of the experimental setup (the ratio of class \mathcal{A}, \mathcal{B} scores). Notably, in a forensic practitioner's methodological validation, probabilistic predictions rather address *proportions of cases* than *error rates* due to the following perspective that has been brought up by forensic experts in various one-on-one conversations: *justice cannot make erroneous decisions* (as it is *just*). Limit Tippett plots provide insights into calibration and discrimination performance without addressing decision costs. As such, the information gain to decision making is indirectly addressed.

For the purpose of examining information performance in greater detail, zoo plots have been introduced in the forensic biometrics com-

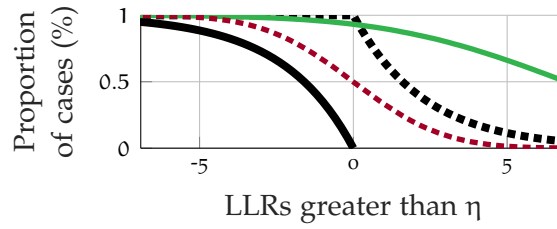


Figure 2.14: Limit Tippett plot example of Fig. 2.3a with universal bounds (black), error rates (red, green), FMR (dashed), and $1 - \text{FNMR}$ (solid).

munity [111–113]. Motivated by *Doddington’s zoo* [114], statistic test visualizations are proposed to classify subjects into four categories based on quantiles in two regards: *hardness to detect* and *easiness to impersonate*. Zoo plots simplify the statistic analysis to quantiles of class \mathcal{A} scores and class \mathcal{B} scores per subject. Fig. 2.15 presents⁵⁴ a zoo plot based on [111, 112]: the zoo distinguishes between so-called *phantoms*, *doves*, *worms*, and *chameleons*, examining the 25% quantiles of class \mathcal{A} and class \mathcal{B} score distributions, which might consider outliers relevant to further algorithmic analyses. Ellipses around the average coordinates per subject illustrate the depending variance estimate assuming Gaussian score distributions. Limit Tippett and zoo plots, however, solely report indirectly on discrimination and calibration performance, in contrast to decision risk and information based measures.

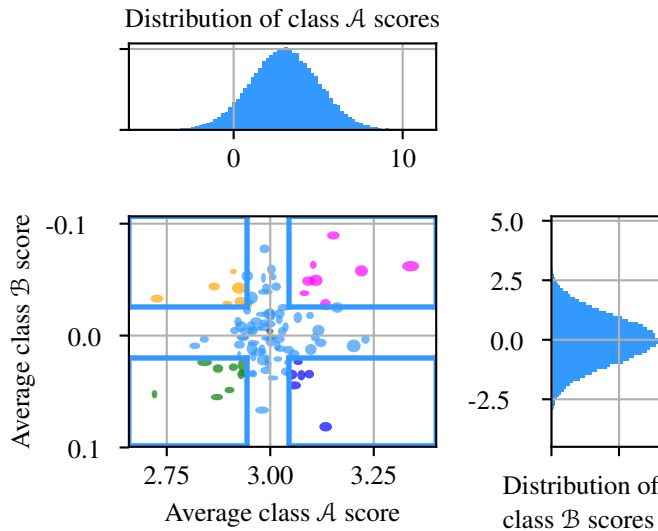


Figure 2.15: Zoo plot example of Fig. 2.3a based on further synthetic reference and probe labels, the zoo addresses phantoms (orange), doves (purple), worms (green), and chameleons (blue), right and on top are the y- and x-axes depending score distributions.

⁵⁴ See: <https://github.com/josbouten/bioplplot>.

2.4.6 Information Performance: System Contribution to Decision Making

Motivated by the Shannon entropy, [27, 115] proposed to estimate the information gain which is yielded by employing a recognition system to decision making instead of reporting error rates. The information gain is reported as the cross-entropy between empirical score distributions and the depending ground-of-truth label distribution. The uncertainty about a set (or partition) \mathcal{U} , see section 2.4.1, is denoted as $H(\mathcal{U})$, i.e., the *entropy of the partitioning* [63, 64]:

$$H(\mathcal{U}) = - \sum_{u \in \mathcal{U}} \Pr(u) \log_2 \Pr(u), \quad (2.44)$$

where in the case of binary partitions with the prior π of one partition:

$$H(\mathcal{U}) = -\pi \log_2 \pi - (1 - \pi) \log_2 (1 - \pi), \quad (2.45)$$

which is known as the *prior entropy* [27, 115]. The *posterior entropy* $H(\mathcal{U} = \{\mathcal{A}, \mathcal{B}\} | S)$ is of concern for measuring the uncertainty of decision making by employing a particular system, informing about evidence by its scores $s \in S$:

$$H(\mathcal{U} = \{\mathcal{A}, \mathcal{B}\} | S) = - \sum_{u \in \mathcal{U}} \Pr(u) \int_{-\infty}^{\infty} \Pr(s | u) \log_2 \Pr(u | s) ds. \quad (2.46)$$

Usually, the computation of Eq. (2.46) is impractical, requiring knowledge of the likelihoods $\Pr(s | \mathcal{A})$ and $\Pr(s | \mathcal{B})$. Systems might be designed to directly provide the ratio of these likelihoods (the LLR), or the likelihoods are unknown, especially when employing discriminative comparators (e.g., Euclidean distance). In [27, 115], the posterior probabilities of a system (denoted by P) are compared with a reference probability distribution (the ground-of-truth labels, denoted by Q). Then, dependencies of the posterior and the likelihood inside the integral are eliminated, and the cross-entropy $H_{Q||P}$ between reference probabilities⁵⁵ Q and system posterior probabilities P is examined:

$$H_{Q||P}(\{\mathcal{A}, \mathcal{B}\} | S) = H_Q(\{\mathcal{A}, \mathcal{B}\} | S) + D_{Q||P}(\{\mathcal{A}, \mathcal{B}\} | S), \quad (2.47)$$

where $H_Q(\{\mathcal{A}, \mathcal{B}\} | S)$ is the posterior entropy of the reference and $D_{Q||P}(\{\mathcal{A}, \mathcal{B}\} | S)$ is the *Kullback-Leibler divergence* [27, 115]:

$$D_{Q||P}(\{\mathcal{A}, \mathcal{B}\} | S) = \sum_{u \in \{\mathcal{A}, \mathcal{B}\}} Q(u) \int_{-\infty}^{\infty} q(s | u) \log_2 \frac{Q(u | s)}{P(u | s)} ds, \quad (2.48)$$

with q denoting the probability density function of Q . The cross-entropy comprises the complementary effects of $H_Q(\{\mathcal{A}, \mathcal{B}\} | S)$, the uncertainty about the propositions if posteriors are computed from the

⁵⁵ The reference (the ground-of-truth) is posterior *as is*.

reference and $D_{Q||P}(\{\mathcal{A}, \mathcal{B}\} | S)$, the divergence of the system posterior P from the reference posterior Q .

By choosing the ground-of-truth labels as reference, i.e., $Q(\mathcal{A} | S) = 1$ if \mathcal{A} is true and $Q(\mathcal{B} | S) = 0$ if \mathcal{B} is true, the entropy of the reference posterior Q is zero: $H_Q(\{\mathcal{A}, \mathcal{B}\} | S) = 0$, such that the cross-entropy becomes the Kullback-Leibler divergence. Supposing the *law of large numbers* holds, Eq. (2.47) (on continuous data) is approximated by the **empirical cross-entropy (ECE)** (on discrete data): $H_{Q||P}(\{\mathcal{A}, \mathcal{B}\} | S) \simeq \text{ECE}$. As the information gain of employing a system within the **Bayesian decision framework (BDF)** is of interest, the depending performance across different prior $Q(u) | u \in \mathcal{U} = \{\mathcal{A}, \mathcal{B}\}$ parameterizations needs to be reported. For binary decision systems, the ECE is computed from class \mathcal{A} scores $S_{\mathcal{A}}$ and class \mathcal{B} scores $S_{\mathcal{B}}$ [27, 115]:

$$\text{ECE} = -\frac{Q(\mathcal{A})}{|S_{\mathcal{A}}|} \sum_{a \in S_{\mathcal{A}}} \log_2 \Pr(\mathcal{A} | a) - \frac{Q(\mathcal{B})}{|S_{\mathcal{B}}|} \sum_{b \in S_{\mathcal{B}}} \log_2 \Pr(\mathcal{B} | b), \quad (2.49)$$

with the posteriors $\Pr(\mathcal{A} | a)$ and $\Pr(\mathcal{B} | b)$. Reformulating Eq. (2.28), the posteriors are expressed by the LLRs $a \in S_{\mathcal{A}}$ and $b \in S_{\mathcal{B}}$:

$$\begin{aligned} \Pr(\mathcal{A} | a) &= \frac{e^a \frac{\pi}{1-\pi}}{1 + e^a \frac{\pi}{1-\pi}} = \left(1 + \frac{1}{e^a \frac{\pi}{1-\pi}}\right)^{-1}, \\ \Pr(\mathcal{B} | b) &= 1 - \Pr(\mathcal{A} | b) = \frac{1}{1 + e^b \frac{\pi}{1-\pi}} = \left(1 + e^b \frac{\pi}{1-\pi}\right)^{-1}, \end{aligned} \quad (2.50)$$

such that Eq. (2.49) is expressed regarding the LLRs a, b :

$$\begin{aligned} \text{ECE} &= \frac{Q(\mathcal{A})}{|S_{\mathcal{A}}|} \sum_{a \in S_{\mathcal{A}}} \log_2 \left(1 + \frac{1}{e^a \frac{\pi}{1-\pi}}\right) \\ &\quad + \frac{1 - Q(\mathcal{B})}{|S_{\mathcal{B}}|} \sum_{b \in S_{\mathcal{B}}} \log_2 \left(1 + e^b \frac{\pi}{1-\pi}\right) \end{aligned} \quad (2.51)$$

and by employing the expectation operator $\langle \cdot \rangle$, the logit function $\text{logit}(x)$ (see Eq. (2.34)), and the sigmoid function $\sigma(x)$ (see Eq. (2.35)):

$$\begin{aligned} \text{ECE} &= \frac{\pi}{\log(2)} \langle -\log \sigma(a + \text{logit}(\pi)) \rangle_{a \in S_{\mathcal{A}}} \\ &\quad + \frac{1 - \pi}{\log(2)} \langle -\log \sigma(-b - \text{logit}(\pi)) \rangle_{b \in S_{\mathcal{B}}}. \end{aligned} \quad (2.52)$$

The ECE can be interpreted as the average information needed to make a decision (between propositions \mathcal{A} and \mathcal{B}) [27, 115]. Thereby, the uncertainty about the propositions is considered (e.g., priors or other knowledge involved in a case).

In (forensic) evaluation, the target prior π is left unspecified by the system designer (e.g., **biometric system vendors** or forensic practitioners), such that systems are compared across all possible π parameter-

izations. Fig. 2.16 illustrates ECE plots⁵⁶ for the systems depicted in Fig. 2.13 depending on the log-odds of the target prior π . Thereby, the **minimum ECE (minECE)** is derived by performing PAV calibration [27, 115], providing the optimal LLR scores. The difference between ECE and minECE measures the information loss due to imperfect score calibration [27, 115].

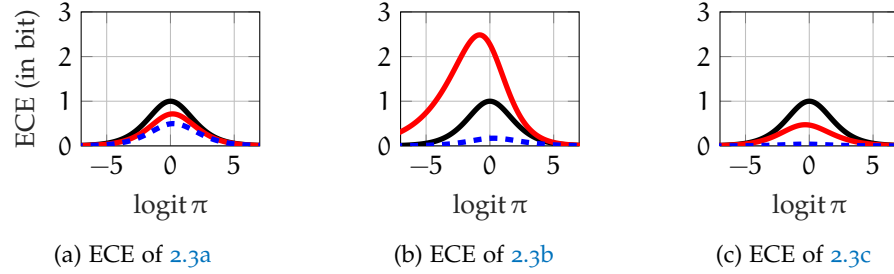


Figure 2.16: Examples of ECE plots with the actual ECE (red), the minECE (blue), and the default performance (black).

In other words, for binary decision systems, PAV resembles the reference (oracle) probability distribution. As the PAV is basis to minECE and also resembles the **ROC's convex hull (ROCCH)** likewise **minimum DCF (minDCF)** [28, 79], a common performance criterion to information and Bayes error rate analyses is foreshadowed.

2.4.7 Goodness of LLR Scores

The *goodness of LLRs*⁵⁷ is referred to as C_{llr} , i.e., the cost of LLRs, which can be derived in different ways by [26–28, 79, 115]:

- integrating over all Bayes risk operating points, likewise the application-independent slope of DCFs:

$$C_{\text{llr}}(S) = \int_0^1 \text{DCF}(S | \tilde{\pi}) d\tilde{\pi}. \quad (2.53)$$

- measuring the accuracy of probabilistic prediction based on **strictly proper scoring rules**⁵⁸, such that proposition-depending

⁵⁶ In [27, 115], the x-axis is referred to regarding \log_{10} odds. However, the natural logarithm is preferred here.

⁵⁷ In the context of the BDF, *goodness* rather reflects the accuracy of scores to fulfill LLR properties, rather than the term *robustness*, usually addressing generalization properties of machine learning. In the context of this dissertation, *goodness* reflects a system's actual contribution to decision making, whereas *robustness* is a subordinated term, which in [24] is referred to as a secondary performance characteristic measured by C_{llr} , **EER**, and the range of LLR values.

⁵⁸ The underlying C_{llr} scoring rule is known as *logistic regression* and is a special case of the canonical form of binary proper scoring rules, parameterized by the Beta

costs are reflected regarding the true set of probabilities (associated with the decision ground-of-truth):

$$\begin{aligned} C_{\text{llr}}(S) &= \frac{1}{2 \log(2)} \left(\sum_{a \in S_{\mathcal{A}}} \frac{\log(1 + e^{-a})}{|S_{\mathcal{A}}|} + \sum_{b \in S_{\mathcal{B}}} \frac{\log(1 + e^b)}{|S_{\mathcal{B}}|} \right) \\ &= \frac{\langle -\log \sigma(a) \rangle_{a \in S_{\mathcal{A}}} + \langle -\log \sigma(-b) \rangle_{b \in S_{\mathcal{B}}}}{2 \log(2)}. \end{aligned} \quad (2.54)$$

- generalizing the ECE at the prior of full uncertainty, i.e., $\pi = 0.5$:

$$C_{\text{llr}} = \text{ECE} |_{\pi=0.5}. \quad (2.55)$$

To sum up, an evaluation setup determines expected LLR values (for its empirical comparisons to be conducted), and C_{llr} examines how good the empirical LLR scores align⁵⁹ yielding (operational) calibration performance, whereas C_{llr}^{\min} assumes well-calibrated scores, i.e., depicting discrimination performance. An intuition to C_{llr} values is established in the speaker recognition community as [28]:

- Systems with $C_{\text{llr}} > 1$ are poorly calibrated, and decisions are better made by omitting these systems or by examining more data in order to reduce the uncertainty of a system provided to decision making.
- Systems with $C_{\text{llr}} = 1$ are as good as a coin toss (on average), i.e., computational effort without gain to decision making.
- Systems with $C_{\text{llr}} < 1$ are *well-calibrated* and aid decision making, regardless of calibration losses are still present or not.

In this dissertation, C_{llr} is the primary evaluation metric summarizing *operational performance* application-independently, alongside with C_{llr}^{\min} resembling the application-independent *discrimination power*.

Notably, the *LLR of the LLR is the LLR*, a premise well-known to forensic statisticians, as an argument put forward by *Niko Brümmer* with a published derivation by *Miranti Indar Mandasari* [39, p. 79]:

1. *Theorem*: scores S are well-calibrated LLRs if the posterior score distribution $\Pr(\mathcal{A} | S)$ is the same as the posterior evidence distribution from the reference–probe comparison $\Pr(\mathcal{A} | E)$.

distribution $\text{Beta}(\alpha, \beta)$ whose parameters α, β (for C_{llr} : $\alpha = \beta = 1$) can be fine-tuned in order to target specific ranges of operating points, such as high-security applications by placing emphasis rather on the left tail of score distributions, e.g., with $\alpha = 2$, providing wider and lower cost minima on higher LLR thresholds [116].

⁵⁹ Thereby, C_{llr} takes the form of the canonical logistic regression loss. Other Bayesian objective functions place emphasis on certain score distribution tails, a practical recipe can be found in [116]. Since C_{llr} also resembles a generalization of these proper scoring rules, this dissertation will refer to C_{llr} as a primary performance characteristic, summarizing all operating points naturally equally.

2. *Corollary:* in other words, the LLRs (left side) encode all comparison information (right side), i.e., the evidence E :

$$\Pr(\mathcal{A} | S) = \Pr(\mathcal{A} | E). \quad (2.56)$$

3. *Corollary:* applying the *logit* transform (log-odds instead of probabilities) and the Bayes' rule to either side:

$$\begin{aligned} \text{logit}(\Pr(\mathcal{A} | S)) &= \text{logit}(\Pr(\mathcal{A} | E)), \\ \log \frac{\Pr(\mathcal{A} | S)}{1 - \Pr(\mathcal{A} | S)} &= \log \frac{\Pr(\mathcal{A} | E)}{1 - \Pr(\mathcal{A} | E)}, \\ \log \frac{\Pr(\mathcal{A} | S)}{\Pr(\mathcal{B} | S)} &= \log \frac{\Pr(\mathcal{A} | E)}{\Pr(\mathcal{B} | E)}, \\ \log \frac{\Pr(S | \mathcal{A}) \frac{\Pr(\mathcal{A})}{\Pr(S)}}{\Pr(S | \mathcal{B}) \frac{\Pr(\mathcal{B})}{\Pr(S)}} &= \log \frac{\Pr(E | \mathcal{A}) \frac{\Pr(\mathcal{A})}{\Pr(E)}}{\Pr(E | \mathcal{B}) \frac{\Pr(\mathcal{B})}{\Pr(E)}}, \\ \log \frac{\Pr(S | \mathcal{A})}{\Pr(S | \mathcal{B})} &= \log \frac{\Pr(E | \mathcal{A})}{\Pr(E | \mathcal{B})}. \end{aligned} \quad (2.57)$$

4. *Proof:* $\log \frac{\Pr(S | \mathcal{A})}{\Pr(S | \mathcal{B})}$ is the LLR of the LLR; $\log \frac{\Pr(E | \mathcal{A})}{\Pr(E | \mathcal{B})}$ is the LLR.

Thus, LLRs are the desired system output and C_{llr} is their measure.

2.5 AUTOMATIC SPEAKER RECOGNITION

Speaker recognition systems⁶⁰ for biometric verification and identification tasks are of similar design: acoustic features are extracted as a voice sample representation in the latent *acoustic* subspace, whereas biometric comparators use projections of these features into the latent *biometric* subspace for the purpose of identity inference. Furthermore, score post processing schemes are employed in order to enhance discrimination and calibration performance. Computationally, state-of-the-art comparators and score post processing schemes are capable of comparing multiple voice references and probe features at once. State-of-the-art technology is suitable for either biometric recognition task.⁶¹ The scope of this dissertation lies on binary decisions, i.e.,

⁶⁰ Parts of this section are based on different collaborative works. The overview of the PLDA comparator family is based on the work with Daniel Ramos and Alicia Lozano-Diez [117]. The overview that aims at non-experts in speaker recognition to proficient readers is derived from the work with Amos Treiber, Jascha Kolberg and Nicholas Evans (among others) [19, 118], a 23 co-author collaboration survey on data privacy in speaker and speech characterization and a follow-up.

⁶¹ Biometric recognition covers two machine learning tasks (verification and identification), other *characterization* tasks involve, e.g., *speaker diarization* (the annotation of the speech sequence with speaker labels that indicate *who spoke when*), searching for speakers, and clustering of voice data. When seeking compliance to the 2017 ISO/IEC standard on harmonized biometric vocabulary [1], the term *biometric characterization* would be a new entry, accounting for a broader sense of *recognition tasks*. This ambiguity is attributable to the different concepts of what *recognition* means.

the verification of biometric identity claims with a binary outcome $\in \{true, false\}$. Research findings of this dissertation, however, might be easily transferable from verification to identification and other tasks, as symmetric and generative comparators are employed following a fully Bayesian paradigm.

The following sections start with a brief overview of speaker recognition technology and how the research field has evolved over time. Then, conventional signal processing methods fundamental to this dissertation are depicted in terms of extracting acoustic and biometric features, comparing voice representations and post processing scores.

2.5.1 Brief Overview of Speaker Recognition Technology

The presentation of this brief overview is aimed at the non-expert reader and so the terminology used below is adapted to that used in other fields of biometrics [1] and computer science. It also covers the full processing pipeline illustrated in Figs. 2.17 and 2.18, which covers both feature extraction and biometric comparison. While the treatment focuses specifically on automatic speaker recognition, many of the techniques described in this section are also applicable to diarization and other speaker characterization applications. Readers proficient in speech signal processing and speaker recognition could continue at section 2.5.3.

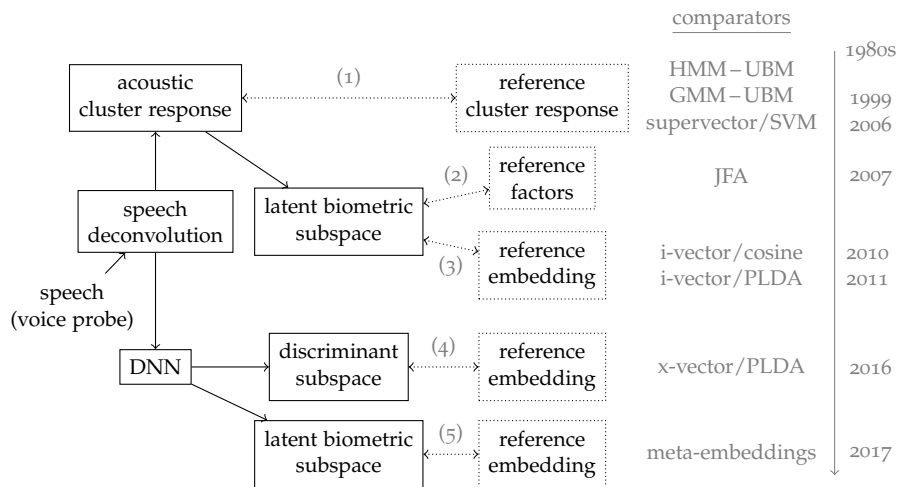


Figure 2.17: Overview of speaker recognition with the typical processing toolchain (solid), groups of comparators (dotted), and timeline.

Fig. 2.17 shows the most dominant speaker recognition technologies. They are all based on the deconvolution of speech signals. This pre-processing step is necessary to separate so-called source and filter components. This is achieved by a process known as homomorphic analysis [119], which is typically applied to short-term intervals of the speech signal. The source component comprises *pitch* and *glottal*

pulse information, whereas the filter component represents *vocal tract* information. Traditionally, acoustic features used for speaker recognition encompass only the latter. In contrast to other fields of biometrics, acoustic features used for speaker recognition are an *inferred* representation of the vocal tract rather than being a directly *observed* biometric characteristic. With speech being a dynamic signal, feature vectors are extracted periodically, e.g., from *sliding windows*, typically representing in the order of 25 ms of consecutive speech and with a 10 ms window offset [120, 121]. While there are a host of alternatives, the most popular acoustic features are *mel-frequency cepstral coefficients* (MFCCs) [122–124].

Approaches to biometric comparison, also shown in Fig. 2.17 by groups of comparators, encompass probabilistic identity inference by subspace identity models and deep neural network (DNN) based feature extraction. Group (1) of Fig. 2.17 corresponds to comparisons using probabilistic cluster responses between an acoustic cluster, i.e., a *universal background model* (UBM) and a reference model, e.g., hidden Markov models (HMMs) [125] or *Gaussian mixture models* (GMMs) [126]. In this case, the biometric information is the reference-specific cluster centroids (the mean values of the probabilistic clusters). The concatenation of these mean values is referred to as a *super-vector* [30]. Groups (2,3) in Fig. 2.17 correspond to techniques that decompose supervectors into factors that represent both biometric and non-biometric subspaces. These factors are referred to as *probabilistic embeddings*, namely latent (inferred rather than observable) representations that lie in a lower dimensional subspace that is more immune to nuisance variation. In group (2), joint factor analysis (JFA) [127–129] explicitly yields *biometric* and *non-biometric* embeddings. JFA scores the likelihood of non-biometric factors given biometric factors as reference. By contrast, group (3) yields a *total variability subspace*, referring to factors as *intermediate-sized vectors* (i-vectors) [130], being *probabilistic embeddings*. Pairs of reference – probe i-vector embeddings are compared by, e.g., cosine distance similarity and *probabilistic linear discriminant analysis* (PLDA) [60, 131–133]. Groups (4,5) encompass *discriminative embeddings* that are derived using DNNs (referred to as *x-vectors*) [134–136] and are also compared to PLDA. In the case of speaker recognition, *biometric information* in the form of *templates* is generally point estimates (usually of high quality), e.g., supervectors, JFA factors, or embeddings (without uncertainty propagation). However, when facing variable quality conditions, uncertainty needs to be propagated in a principled manner, e.g., i-vector embeddings estimated alongside their uncertainty are a *model* rather than a *tem-*

plate.⁶² Group (5) relates to the extraction of *meta-embeddings* [54, 99], which are capable of principled uncertainty propagation.

In groups (1-5), *biometric information* is represented by *models*. Algorithmically, templates are expected values of some form (averages of high precision), whereas models center data (remove averages of low precision) to propagate uncertainties on the biometric similarity versus on the biometric dissimilarity. Algebraically, groups (1-4) rely on *logsums* and *inner products* (*dot products*), whereas group (5) and end-to-end uncertainty propagation also rely on *matrix inversions* and *log-determinants*. As such, computations are carried out on *floating point* representations as opposed to *integer* values.

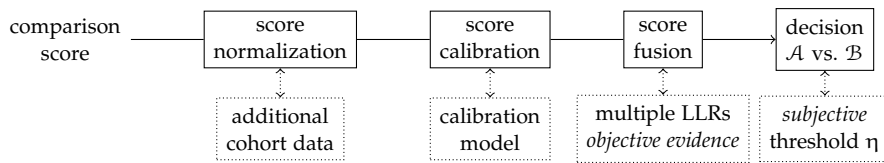


Figure 2.18: Overview on score processing for the purpose of making good informed decisions on average.

Whatever the approach to speaker recognition, some form of normalisation [30] is usually applied to compensate for nuisance variation, e.g., cepstral mean and variance normalization that marginalizes microphone and other channel effects via normally distributed data. Other reasons to normalize scores are to improve system calibration and fusion. A general approach to score post-processing is illustrated in Fig. 2.18. A large, auxiliary set of cohort data is often used in conjunction with references and probes to normalize the scores produced by some of the systems described above. Score calibration is often applied to transform scores into *log-likelihood ratios* (LLRs), whereas score fusion techniques can be applied to improve reliability by combining the scores produced by different speaker recognition systems, e.g., so that they produce LLRs that reflect the weight of evidence for a given probe in a given reference–probe comparison. The use of LLRs rather than raw scores has distinct advantages, e.g., (a) the Bayes decision risk is minimized, and (b) decision making is solely inferred from biometric information, namely to the proportion of *mated* and *non-mated* reference–probe pairs (ideally, encoded by a score *in its value*: the LLR).

This dissertation relates to the state-of-the-art as follows. The i-vector/PLDA paradigm (group 3) is the baseline. The benefit of this paradigm over deep learning systems is the inherent data interpretability throughout each processing step; data interpretability is di-

⁶² The vast majority of literature assumes high precision after embedding extraction (for the sake of computational effort), such that embeddings are treated as templates (approximated from models) and the uncertainty of the feature estimate is not considered further. By contrast, [137] propagates the uncertainty of i-vector embeddings (as models) in a principled manner throughout the PLDA comparison.

luted during (current) DNN processing, DNNs are thus inapplicable to scenarios which demand understanding of each step in evidence reporting. In this dissertation, *quality vectors (q-vectors)* are proposed—quality estimates derived from acoustic i-vectors—and employed in score normalization and calibration to sustain good decision making in unconstrained environments. In other words, the data flow of a speaker recognition system is *exploited* during processing—data representations of each step have a well-defined interpretation—but the baseline system itself remains *unchanged*. This benefits *biometric system operators* who use and understand state-of-the-art technology but are not developing speaker recognition systems end-to-end. Using the methods proposed in this dissertation, operators are capable of not only employing state-of-the-art technology but moreover of gaining the ability to utilize licensed systems in *unconstrained environments*. Based on the *Bayesian decision framework (BDF)* score normalization/calibration is nothing but threshold normalization/calibration throughout—an adaptation of the belief in a decision policy facing new information (e.g., about quality) is formalized in the same way as an update of the strength of evidence (LLR scores) in the light of the same new information.

In the following, the extraction of acoustic features is briefly outlined. Early biometric comparators in speaker recognition are directly employed on these acoustic features (as well as the latest end-to-end DNNs). The ideas and concepts of conventional comparators are presented for two purposes. For one, their chronological evolution is fundamental to the extraction of biometric from acoustic features. For another, to provide a brief outline of involved concepts for computer scientists (being non-experts in this field) that might find struggle with (although well-written) tutorial overviews on speaker recognition (e.g., [30, 31]).

2.5.2 Acoustic Feature Extraction in Brief

The survey [30] distinguishes between the following acoustic features:

- **High-level.** Features represent behavioral characteristics, such as accent, idiolect, or phones, typically examined by linguists and phoneticians in a semi- or non-automated fashion.
- **Prosodic and spectro-temporal.** Features comprise behavioral and physiological information, e.g., rhythm, pitch, or harmonics, which are also influenced by the properties of a biometric subject's vocal tract, where research contributions are coming from signal processing and acoustic phonetics.
- **Short-term spectral and voice source.** Features rather put emphasis on physiological characteristics, such as length and dimension of the vocal tract, which are indirectly examined uti-

lizing spectral and other easily extractable low-level signal features, enabling automated real time recognition.

Short-term spectral and voice source features are the most chosen features for automatic speech processing. In speaker recognition, acoustic features are extracted from voice samples in order to assess the biometric evidence [31]. Typically employed acoustic features are, e.g., MFCCs, summarizing a perceptual scale of pitches [123, 124]; linear predictive coding coefficients (LPCCs), a filter-source model of speech resonances [138, 139]; and relative spectral transform and perceptual linear predictions (RASTA-PLPs), smoothed spectra warping robust to linear spectral distortions [140–142]. Among others, the speaker recognition community has introduced: *Perseus* features, more noise robust variation of MFCCs with responses from Gammatone filter bank [143], and constant Q cepstral coefficients (CQCCs), geometrically distributed octaves while preserving linearly spaced frequency filters [144]. Effectively, these features share one property, exemplarily presented in the following: acoustic features are extracted from a homomorphic transform of speech for the purpose of speech deconvolution [119].

Example: Deconvolution of Speech, a Homomorphic Analysis

Canonically, a time t dependent signal $u(t)$ is deconvoluted by a linear filter system into its (e.g., two) components $u_1(t), u_2(t)$. Let \otimes denote the convolution of speech in the observation space, \oplus denote the convolution in the homomorphic space (preserving algebraic structures of speech), and ϕ the homomorphic transform of the linear filter system, such that the structure of speech convolution is preserved in the homomorphic space—the homomorphic transform of a convoluted signal equals the homomorphic convolution of its transformed components:

$$\phi(u(t)) = \phi(u_1(t) \otimes u_2(t)) = \phi(u_1(t)) \oplus \phi(u_2(t)). \quad (2.58)$$

Thereof, speech components are easier to deconvolute in the homomorphic space.

According to [119] (published in 1968; concurrent nevertheless), there are many equivalent representations of such filter systems in speech processing, where the most straightforward and most generally applicable are based on the complex Fourier transform. The excitation $u(t)$ and the response $\hat{u}(t)$ of the system are related by their associated complex Fourier transforms $U(j\omega), \hat{U}(j\omega)$ as:

$$\hat{U}(j\omega) = \log U(j\omega) = \log |U(j\omega)| + j\Phi(j\omega), \quad (2.59)$$

with the complex phase $\Phi(j\omega)$ associated with $U(j\omega)$; denoting complex numbers in terms of radians of the angle ω with $-\pi < \omega < \pi$, imposing $j^2 = -1$. By applying the inverse Fourier transform, the transform of the linear filter system is realized, and

acoustic features are extracted. As the Fourier transform is transforming orthogonally, it results in *decorrelated* Fourier coefficients.

By equivalently substituting the Fourier transform with the *Z-transform*, the system is evaluated on the unit circle. The Z-transform of the time-discrete signal $u(n)$ (with integer indices $n \in \mathbb{Z}$) is the *formal power series* $U(z)$ with the complex number z :

$$U(z) = \sum_{n=-\infty}^{\infty} u(n) z^{-n}. \quad (2.60)$$

The *input sequence* (the filter input) and the *complex cepstrum* (the filter response; an anagram of the processed *spectrum* $U(z)$) are represented as $u(n)$, $\hat{u}(n)$ with the associated Z-transforms $U(z)$, $\hat{U}(z)$, such that $\hat{U}(z) = \log U(z)$. The *inverse Z-transform* reveals the complex cepstrum $\hat{u}(n)$ in terms of its additive components:

$$\hat{u}(n) = \frac{1}{2\pi j} \oint \log U(z) z^{n-1} dz. \quad (2.61)$$

Speech is considered to be a convolution of *pitch* $p(n)$, *glottal pulse* $g(n)$, and *vocal tract* $v(n)$ sequence representations [119], where the speech waveform is viewed through a window $w(n)$, such that:

$$u(n) = [p(n) \otimes g(n) \otimes v(n)] w(n). \quad (2.62)$$

The complex cepstrum is the basis of speech signal processing.

The convolved speech components are transformed into additive components in the complex cepstrum, making them separable in an orthogonal acoustic space—ideally, cepstral representations are decorrelated. Depending on the chosen feature type, amplitude and/or phase, information is used (e.g., MFCCs ignore the complex phase and just use amplitude information). Furthermore [30], [voice activity detection \(VAD\)](#) (removal of non-speech), *speech enhancement* (signal denoising), and *feature normalization* methods can be applied (to fulfill the requirements of following signal processing, such as zero mean and unit variance).

In this dissertation, solely short-term spectral acoustic features are used, namely MFCCs. To explicitly define the used acoustic features [31]: MFCCs are obtained after a short-term Fourier transform of 25 ms voice segments⁶³, which are sampled by *sliding windows* [120, 121] at a rate of 100 Hz, resulting in the *amplitude* and *phase* contribution of the examined frequency bands.⁶⁴ Then, the power spectrum is computed, filtered by the logarithmic *mel filterbank*, and *cepstral coefficients* are obtained by a discrete cosine transform of the logarith-

⁶³ In the [automatic speech recognition \(ASR\)](#) and speaker recognition communities usually referred to as *frames* [120, 121].

⁶⁴ Considering telephone speech (8 kHz denotes the upper frequency bound of the communication channel) usually, e.g., 200 Hz to 3.8 kHz frequencies are examined.

mic values of each filterbank energies. Signal phase information is thereby omitted, solely considering amplitude information. A speech segment representation of compressed energy information results.

In general, acoustic features can be augmented and/or normalized [30, 31] by, e.g., cepstral mean and variance normalization, applying *feature warping* [145] or *RASTA filtering* [142]. Typically, *static* cepstral coefficients are augmented by first and second order derivatives [120, 121], i.e., the velocity and acceleration of cepstral coefficients, which are stacked onto the static features, forming the representation of a short-term voice segment.⁶⁵ In this dissertation, 60 acoustic features are extracted from 20 static MFCCs.

MFCCs are the basis for conventional signal processing, the estimation of latent biometric subspaces, and end-to-end DNN approaches. In the following, conventional signal processing and comparison methods are outlined, forming the theoretical basis of the state-of-the-art technology employed in this work.

2.5.3 Conventional Signal Processing and Comparison

Conventional signal processing methods in speaker recognition rely on so-called **universal background models (UBMs)**, cf. groups (1-3) of Fig. 2.17. In terms of Fig. 2.17, the response of an acoustic cluster (referred to as the UBM) with voice sample representations as input describes these samples in a known acoustic scenario. UBMs are statistic clusters (e.g., HMMs and GMMs), modeling the space of observable acoustic features. Depending on the employed comparison algorithm, UBMs serve different purposes: in *HMM-UBM*, *GMM-UBM*, and *supervector* comparisons, group (1) of Fig. 2.17, speaker-dependent models are adapted from a UBM modeling proposition \mathcal{A} , whereas the UBM serves as the model to proposition \mathcal{B} during **LLR** computation. However, for the purpose of extracting *latent factors*, group (2), or *probabilistic embeddings*, group (3), the UBM is basis to the estimation of a latent voice sample representation, i.e., to the sample representation embedded in a probabilistic subspace of the UBM. The latter is crucial for the extraction of i-vector features.

UBMs as (state-transitioning) HMMs [125] model temporal speech patterns by transitioning between latent states capable of emitting feature vectors (text-dependent models). UBMs as (state-less) GMMs [126] cluster speech patterns probabilistically (text-independent models). GMMs can be seen as single-state HMMs that are also being capable of emitting feature vectors. GMMs $\lambda = \{w_c, \mu_c, \Sigma_c \mid c \in C\}$ are probabilistic clusters and compute log-likelihoods by a weighted log-sum over their clusters $c \in C$ (referred to as *mixture components*) with

⁶⁵ Conventionally, 13 or 20 static MFCCs are used (replacing the zeroth MFCC with a frame's energy). When employing first and second order derivatives, 39 or 60 features are yielded per voice segment.

weights w_c , component-wise normal distributions $\mathcal{N}(\boldsymbol{\psi}_t | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ having distribution mean and covariance parameters $\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c$ [31]:

$$\log \Pr(\boldsymbol{\psi}_t | \boldsymbol{\lambda}) = \log \sum_{c \in \mathcal{C}} w_c \mathcal{N}(\boldsymbol{\psi}_t | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c). \quad (2.63)$$

Thereby, the mean values represent cluster centroids, and covariance terms represent the precision of a component (precision $\boldsymbol{\Sigma}_c^{-1}$ being the inverse of uncertainty $\boldsymbol{\Sigma}_c$): small covariances indicate high precision, such that solely features close to a centroid yield high likelihoods and vice versa—high covariances indicate poor precision. LLRs are computed by comparing the likelihoods of a speaker GMM to a UBM. Log-likelihoods $\log \mathcal{N}(\boldsymbol{\psi}_t | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ are estimated per GMM component λ_c . Notably, probability density functions (pdfs) $f_X(\mathbf{x} | \boldsymbol{\lambda})$ of the exponential family, such as the multivariate normal distribution $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ with mean and covariance *expectation parameters* $\boldsymbol{\mu}, \boldsymbol{\Sigma}$, can be expressed by their *natural parameters* $\boldsymbol{\eta}(\boldsymbol{\lambda})$, outlining the distribution, and by *sufficient statistics* $\mathbf{T}(\mathbf{x})$, describing the data for that distribution [59]:

$$f_X(\mathbf{x} | \boldsymbol{\lambda}) = \exp \left(\boldsymbol{\eta}(\boldsymbol{\lambda})^T \mathbf{T}(\mathbf{x}) + A(\boldsymbol{\lambda}) + B(\mathbf{x}) \right) \quad (2.64)$$

with distribution and data-dependent normalization terms $A(\boldsymbol{\lambda}), B(\mathbf{x})$. For a data series $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ with N data samples, the pdf $f_X(\mathbf{X} | \boldsymbol{\lambda})$ is denotable as:

$$f_X(\mathbf{X} | \boldsymbol{\lambda}) = \exp \left(\boldsymbol{\eta}(\boldsymbol{\lambda})^T \sum_{n \in \mathcal{N}} \mathbf{T}(\mathbf{x}_n) + N A(\boldsymbol{\lambda}) + \sum_{n \in \mathcal{N}} B(\mathbf{x}_n) \right). \quad (2.65)$$

These normalization terms sustain a fundamental probability property: the area under the pdf equals one ($\int f_X(\mathbf{x} | \boldsymbol{\lambda}) d\mathbf{x} = 1$). In other words, log-likelihoods $\log f_X(\mathbf{x} | \boldsymbol{\lambda})$ are computable by linear algebra. For this dissertation, the main benefit of this pdf formulation is not to investigate different members of the exponential distribution family but rather to outline the necessary computations in which distribution pre-calculations are separated from data depending calculations, e.g., occurring during a verification attempt. Consequently, computations are carried out efficiently and for privacy centered communication, and multi-party LLR computation protocols are definable at the precise instances of model parameter and biometric depending data calculations.

For multivariate Gaussian distributions with feature dimension F , the well-known pdf is expressed in terms of the inner product (the dot-product) between the *natural parameters* $\boldsymbol{\eta}(\lambda_c)$ for a GMM component λ_c and the sufficient statistics $\mathbf{T}(\boldsymbol{\psi}_t)$ of acoustic features $\boldsymbol{\psi}_t$. For one feature $\boldsymbol{\psi}_t$ and for a feature sequence $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_T$, the observa-

tion log-likelihoods $\log \mathcal{N}(\psi_t | \mu_c, \Sigma_c)$, $\log \mathcal{N}(\psi_1, \dots, \psi_T | \mu_c, \Sigma_c)$ are (with summarized normalization terms) [59]:

$$\begin{aligned} \log \mathcal{N}(\psi_t | \mu_c, \Sigma_c) &= \left(\eta(\lambda_c)^T \mathbf{T}(\psi_t) \right) \\ &\quad - \left(\frac{1}{2} \mu_c^T \Sigma_c^{-1} \mu_c + \frac{1}{2} |\Sigma_c| + \frac{F}{2} \log(2\pi) \right), \\ \log \mathcal{N}(\psi_1, \dots, \psi_T | \mu_c, \Sigma_c) &= \left(\eta(\lambda_c)^T \sum_{t \in T} \mathbf{T}(\psi_t) \right) \\ &\quad - T \left(\frac{1}{2} \mu_c^T \Sigma_c^{-1} \mu_c + \frac{1}{2} |\Sigma_c| + \frac{F}{2} \log(2\pi) \right) \\ \text{with} \quad \eta(\lambda_c) &= \begin{bmatrix} \Sigma_c^{-1} \mu_c \\ -\frac{1}{2} \Sigma_c^{-1} \end{bmatrix} \quad \text{and} \quad \mathbf{T}(\psi_t) = \begin{bmatrix} \psi_t \\ \psi_t \psi_t^T \end{bmatrix}. \end{aligned} \quad (2.66)$$

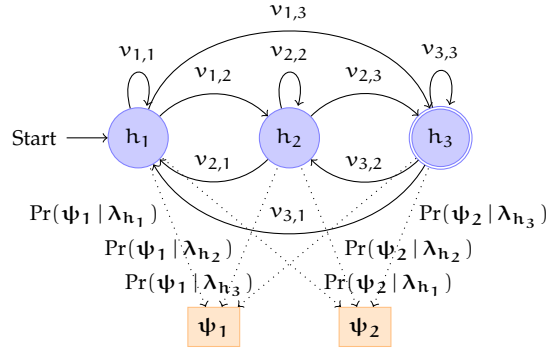
In Eq. (2.66), T denotes the transposition operator, $t \in T$ time of the feature vectors and $\mathbf{T}(\psi_t)$ the sufficient statistics (disambiguation by font style). Notably, for an entire voice sample representation $\Psi = \{\psi_1, \dots, \psi_T\}$, the sufficient statistics are the sole data dependent term and are accumulated before computing the inner product. Terms depending on F, μ_c, Σ_c are pre-computable constants. Effectively, log-likelihoods of a Gaussian distribution are computed by addition and multiplication, where non-linear terms (such as $\log(2\pi)$) are precomputable constants.

UBMs are trained by the iterative [expectation-maximization algorithm \(EM\)](#) [59], optimizing the model fit of the training data: during expectation steps, the *expected sufficient statistics* are computed. Afterwards, during maximization steps, the log-likelihoods of the entire training data set are optimized w.r.t. the model parameters, see [59, 120, 121] for details. In speaker recognition (e.g., [146]), UBM training is *hierarchical*. Therein, the first EM iterations optimize the fit of a single cluster (UBM component), which is then split into two depending on the principal axis of variance [146, source code]. Then, the fit of two clusters is optimized during consecutive EM iterations, thereafter of four clusters and so on, until the final number of clusters, i.e., the final number of UBM components is—usually—a power of two.⁶⁶ During enrolment, the UBM is seen as the prior distribution and the enrolment data as new information, such that the model is adapted to the *maximum a-posterior (MAP)* model, see [59, 120, 121].

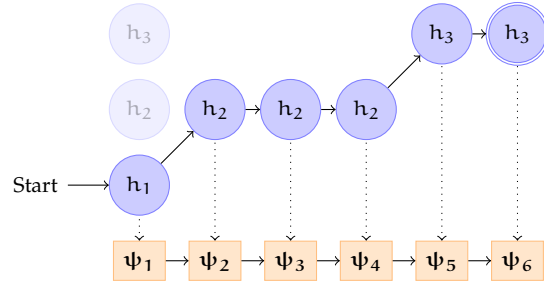
In the following examples, the concepts of the group (1) comparators in Fig. 2.17 are briefly outlined, providing a background on the history of research in the speaker recognition community. Chronologically by appearance in the literature, the HMM–UBM comparator is depicted first, followed by the GMM–UBM comparator. Afterwards,

⁶⁶ Clusters of too low or high precision might be omitted to neglect acoustic artifacts of the training data or components of poor fit. When omitting components, the remaining UBM weights need to be rescaled, such that their sum is one.

supervector comparison by SVMs is described for one of the many variants proposed in the literature pool.



(a) Observing two feature vectors ψ_1, ψ_2 , each state with its conditional likelihood of the form $\Pr(\psi | \lambda)$ and transition probabilities of the form v .



(b) Observing six feature vectors ψ_1, \dots, ψ_6 and an (exemplarily) most likely path; opaque states indicate each state's location in the transition graph.

Figure 2.19: HMM example: finite state machine with three states h_1, h_2, h_3 (blue circles) with h_1, h_3 as start and final states, observing feature vectors (orange boxes); transition probabilities (solid) and emission probabilities (dotted).

Example: Acoustic Features, the HMM–UBM Comparator

HMMs [125] are finite state machines with underlying statistical models. They describe a sequence of possible (latent) states by probabilities based on the previously attained sequence of states. The transition likelihood between two states is modeled by transition probabilities $V = \{v_{i,j} \in \mathbb{R} \mid 1 \leq i, j \leq k, 0 \leq v_{i,j} \leq 1, \sum_i v_{i,j} = 1\}$ with k HMM states, and the sum of transition probabilities of exiting a state equals one. Fig. 2.19 illustrates an exemplary HMM as a graphical model when observing data.

Each state $H = \{h_1, \dots, h_k\}$ is capable of estimating likelihoods to data observations, as each state resembles a latent variable outlined by a mixture λ of C (e.g., Gaussian) distribution components. The probability $\Pr(\psi_t | \lambda_h)$ of an HMM state h observing data ψ_t with

time-dependent frame index $t \in T$ and T feature vectors from a mixture of multivariate Gaussian distributions \mathcal{N} is:

$$\Pr(\psi_t | \lambda_h) = \sum_{c \in C} w_c \mathcal{N}(\psi_t | \mu_c, \Sigma_c), \quad (2.67)$$

where w_c, μ_c, Σ_c denote component weights, mean vectors and covariance matrices (per state). Notably, as acoustic features are assumed to be decorrelated (the feature vector elements are independent), covariances matrices are diagonal. States are capable of either associating a likelihood to any observed data (data discrimination) and to emit any observable data for a given likelihood (data generation). The most likely latent state transition path is determined via *forward-backward algorithms*, e.g., by the *Viterbi algorithm* or by the *Baum-Welch algorithm* [125]. Considering state transition and observation/emission probabilities, HMMs are capable of discriminating any observed data sequence, with likelihoods depending on the assumed sequence of latent states and generating (emitting) any data sequence for given state sequences and emission likelihoods.

The outline of HMMs is determined by their number of states, state transitions and each state's distributions, e.g., when modeling each phoneme (each sound unit) by another HMM, each HMM requires less states to sufficiently represent the data, fewer state transitions and fewer mixture components per state. Usually, speaker-dependent HMMs are adapted from the UBM based on few enrolment voice samples. In consequence, of the limited availability of enrolment data⁶⁷, not all UBM parameters are adapted towards a speaker's subspace, such that transition, weight, and covariance parameters of the UBM are assumed to be equal among all subjects. Conventionally, HMMs model passphrases, words, or phonemes, making them suitable for text-dependent application scenarios. For text-independent systems, however, GMMs are more suitable.

Example: Acoustic Features, the GMM – UBM Comparator

GMMs [126] are an HMM special case with one state (without state transitions, thus effectively state-less), where likelihoods are computed from a mixture of Gaussian distributions, see Eq. (2.63). Similarly to the HMM – UBM approach, a UBM (a universal GMM) serves in the GMM – UBM comparison as the proposition \mathcal{B} model, whereas a speaker-dependent GMM is adapted from the UBM during enrolment by using the enrolment data of a subject, serving as the proposition \mathcal{A} model.

Usually, only GMM's mean values μ_c are updated, assuming equal weights w_c and covariances Σ_c among all subjects for com-

⁶⁷ Hours of uttered speech would aid the sufficiency of training models. To utter these speech amounts, however, could be inconvenient to end-users.

ponents $c \in C$ (due to limited enrolment data). As HMMs usually employ $C = \{8, 16, 32\}$ components per state, speaker recognition GMMs often comprise $C = \{512, 1024, 2048, 4096\}$ components, leading for 60-dimensional acoustic features to up to 14 995 456 free parameters to estimate in terms of weights, means, and covariances (495 616 free parameters for diagonal covariance approximations with 245 760 speaker-depending mean values).

In contrast to HMMs, GMMs cannot model state transitions but are, in turn, better suited to efficient computations in text-independent application scenarios. However, as the GMM mean values of reference models solely represent a speaker's identity as a biometric subject, the research focus of the speaker recognition community moved to the so-called *supervectors*.

Example: Acoustic Features, Supervector/SVM Comparator

The concatenation of mean values across mixture components is referred to as *supervector* $\mathbf{m}^T = [\mu_1^T, \dots, \mu_C^T]$ [30]. The UBM parameters w_c, Σ_c remain speaker-independent as non-biometric information. Adapted reference and probe supervectors are compared by SVMs in terms of the divergence between both clusters represented [147]. Supervectors are high-dimensional representations of voice samples, employable to biometric and non-biometric recognition tasks (the UBM models the acoustic subspace, not the biometric subspace). Thereof, supervectors are considered as *sparse data*, since reference and probe samples usually convey a few seconds up to some minutes of speech—the linguistic variability modeled by text-independent UBMs is not representative within single voice samples.

Speaker recognition based on supervectors is carried out by employing discriminative models (e.g., SVMs) or generative models (e.g., probabilistic embeddings). In contrast to conventional SVM kernels, the *GMM supervector kernel* [147] computes a bound to the *Kullback-Leibler divergence* between two GMM distributions, one represented by the reference supervector \mathbf{m}_r , another by the probe supervector \mathbf{m}_p . By assuming diagonal covariances, a closed-form solution is represented by its corresponding inner product, i.e., the kernel function $K(\mathbf{m}_r, \mathbf{m}_p)$:

$$K(\mathbf{m}_r, \mathbf{m}_p) = \sum_{c \in C} \left(\sqrt{w_c} \Sigma_c^{-\frac{1}{2}} \mathbf{m}_{c,r} \right)^T \left(\sqrt{w_c} \Sigma_c^{-\frac{1}{2}} \mathbf{m}_{c,p} \right), \quad (2.68)$$

where $\mathbf{m}_{c,r}, \mathbf{m}_{c,p}$ represent component-depending terms of the reference and probe supervector. A two-class SVM is constructed from sums of the kernel function with L support vectors (denoted by \mathbf{m}_s), i.e., supervectors outlining the SVM's decision bound that are

determined by a training process. Discriminative scores S_{SVM} are computed by speaker-depending SVMs as:

$$S_{\text{SVM}}(\mathbf{m}_p) = \sum_{l \in L} \alpha_l \theta_l K(\mathbf{m}_{s,l}, \mathbf{m}_p) + b, \quad (2.69)$$

where θ_l are the ideal outputs associated with the support vectors (0, 1 class labels for class \mathcal{B} and class \mathcal{A}), consisting of non-/mated and mated supervectors from the reference and the background, and positive weights α_l sustain the condition: $\sum_{l \in L} \alpha_l \theta_l = 1$.

The outline of SVMs regarding supervectors is a straight-forward consequence from the assumptions on GMM–UBM comparisons, i.e., equal covariance terms and equal component weights across all subjects (due to insufficient data in enrolment and recognition). As solely supervectors represent a sample of a speaker in the GMM–UBM approach, comparators such as SVMs that assess these voice representations are designed. Regarding biometric information in the form of templates and models, SVMs are designed to treat features as templates.

Elaborating on noise robustness for SVMs in speaker recognition, [147] proposes *nuisance attribute projection*, removing the sample but not subject depending subspace from supervectors by projection. In [148], *within class covariance normalization* (WCCN) is proposed, a supervised *whitening* transform: the feature space is projected into a more discriminant subspace, where feature elements are decorrelated (covariances of feature elements are zero), and each feature element's variance is one. The transformation matrix is estimated by the *expected within-covariance class over all classes* [148]. For GMM supervector SVMs, an SVM is required per speaker, distinguishing between the subject's identity and others.

By contrast, other comparators allow to simultaneously compute comparison scores for multiple subjects. Depending on the kernel design, however, comparison can accommodate for features as models. Nevertheless, in the context of this work, templates cannot be used to predict other sample representations of a subject, whereas models can.⁶⁸ This is relevant when comparators trained on idealistic

⁶⁸ Here, the central aspect of models are the formulation of the *posterior marginal* and the *posterior predictive*. For details see [59]. The likelihood $\Pr(\mathcal{D}|\boldsymbol{\eta})$ of a data set \mathcal{D} depends on the natural parameters $\boldsymbol{\eta}$. By modeling priors of the natural parameters $\boldsymbol{\eta}$ and of the data $\Pr(\boldsymbol{\eta}), \Pr(\mathcal{D})$, likelihoods are used to compute the posterior of the natural parameters given the data $\Pr(\boldsymbol{\eta}|\mathcal{D}) = \Pr(\mathcal{D}|\boldsymbol{\eta}) \frac{\Pr(\boldsymbol{\eta})}{\Pr(\mathcal{D})}$. Let the natural parameters exemplarily be based on two terms like the data mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, then the posterior of the natural parameters equals the posterior of these terms $\Pr(\boldsymbol{\eta}|\mathcal{D}) = \Pr(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D})$ since $\boldsymbol{\eta}$ is a function of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. By using *marginalization*, the prior belief in these parameters is updated: the *posterior marginal* of the covariance is $\Pr(\boldsymbol{\Sigma}|\mathcal{D}) = \int \Pr(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D}) d\boldsymbol{\mu}$ and the *posterior marginal* of the mean is $\Pr(\boldsymbol{\mu}|\mathcal{D}) = \int \Pr(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D}) d\boldsymbol{\Sigma}$. The *posterior predictive* $\Pr(\mathbf{x}|\mathcal{D})$ uses the updated belief in a distribution's parameters to predict data \mathbf{x} given a data set \mathcal{D} in a meaningful

conditions are facing environments of unconstrained quality. Therefore, embeddings are extracted and compared by models generalizing across subjects.

2.5.4 Embeddings: Feature Extraction

In this dissertation, *embeddings* are subspace representations w.r.t. a recognition task in which the terms *latent* (or *hidden*) refer to these embeddings as *variables* living in this subspace. Embeddings are estimated probabilistically (e.g., as *biometric factors* [127] and as *i-vectors* [130]) or by DNNs (e.g., as *x-vectors* [134, 135]). In speaker recognition, comparisons of reference and probe embeddings are carried out either in terms of correlation (e.g., by the cosine distance similarity) or in terms of subspace identity models (e.g., by PLDA [60]). Motivated by *stochastic variational Bayes*, so-called *meta-embeddings* [54, 99] propose an identity inference scheme, combining the advantages of probabilistic and discriminative embeddings.

2.5.4.1 Probabilistic Embeddings

Groups (2,3) of Fig. 2.17 examine the biometric subspace of a voice sample *within* an acoustic cluster: emphasis is put on factors representing the biometric voice by decomposing supervectors \mathbf{m} into biometric and non-biometric components. In JFA, group (2), sparse supervectors (with dimensions between 20 k and 250 k) are decomposed into lower-dimensional non-sparse factors (e.g., with 50 to 250 dimensions) representing biometric \mathbf{y} , non-biometric \mathbf{x} , and residual \mathbf{z} latent variables [127, 128].

Example: Latent Biometric Factors, JFA Comparison

JFA hyperparameters $\mathbf{V}, \mathbf{U}, \mathbf{D}$ are pre-trained, mapping supervector components into latent factors $\mathbf{y}, \mathbf{x}, \mathbf{z}$ (or *probabilistic embeddings*). The JFA model is defined regarding the UBM supervector \mathbf{m}_{UBM} (as a data offset) and is expressible in different ways:

manner—the likelihood of the predicted data is based on the knowledge of a data set. In terms of a ratio of marginal likelihoods, the posterior predictive can be easily evaluated as $\Pr(\mathbf{x}|\mathcal{D}) = \frac{\Pr(\mathbf{x}, \mathcal{D})}{\Pr(\mathcal{D})} = \int \int \Pr(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\boldsymbol{\mu} d\boldsymbol{\Sigma}$ [59].

On the one hand, the posterior predictive is usable to estimate (train) likelihood models, e.g., speaker dependent HMMs and GMMs (to enrol reference models). On the other hand, by formulating latent variables rather than distribution parameters, latent class subspaces can be estimated. For example, they can be used to recognize, e.g., a language, an emotion, and a biometric subject. In this dissertation, these concepts are used to estimate latent variables (probabilistic embeddings) to represent speaker identities based on a formal inference from acoustic voice data. Note: despite the appearance of Bayesian statistics, most computations involved are nothing but addition and multiplication (the most used distributions have elegant *closed-form solutions*).

$$\begin{aligned}
\mathbf{m} &= \mathbf{m}_{\text{UBM}} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \mathbf{D}\mathbf{z} \\
\Leftrightarrow \quad \mathbf{V}\mathbf{y} &= (\mathbf{m} - \mathbf{m}_{\text{UBM}}) - (\mathbf{U}\mathbf{x} + \mathbf{D}\mathbf{z}) \\
\Leftrightarrow \quad \mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z} &= \mathbf{m} - (\mathbf{m}_{\text{UBM}} + \mathbf{U}\mathbf{x}) \\
\Leftrightarrow \quad \mathbf{U}\mathbf{x} &= (\mathbf{m} - \mathbf{m}_{\text{UBM}}) - (\mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z}). \quad (2.70)
\end{aligned}$$

The braces foreshadow the concept of how JFA is used: (i) the JFA model decomposes supervectors; (ii) having multiple [enrolment](#) samples, biometric supervector components $\mathbf{V}\mathbf{y}$ are estimated from all centered supervectors of reference samples $\mathbf{m} - \mathbf{m}_{\text{UBM}}$, thereafter, sample-depending non-biometric supervector components are estimated with the residuals $\mathbf{U}\mathbf{x} + \mathbf{D}\mathbf{z}$. Therein, the speaker-depending biometric and residual factors (\mathbf{y}, \mathbf{z}) form the reference model (iii) of a biometric subject. The depending term in the supervector space $(\mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z})$ is derived by jointly centering the supervector mean by the UBM supervector and the sample dependent non-biometric component term $\mathbf{U}\mathbf{x}$. During recognition (iv), non-biometric supervector components $\mathbf{U}\mathbf{x}$ of the voice probe are decomposed by the centered supervector term $\mathbf{m} - \mathbf{m}_{\text{UBM}}$ that is re-biased by the joint biometric and residual term $\mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z}$ (the enrolled reference).

Conceptually, supervectors \mathbf{m} are obtained from iteratively adapting the UBM, where the adaptation depends on the UBM's cluster response regarding acoustic features ψ_1, \dots, ψ_T . Notably, the UBM adaptation is deterministic, such that its response (of the first adaptation iteration) outlines the estimation of supervector decomposition. Thus, the JFA hyperparameters are trained to fit the decomposition model accordingly. Therefore, a UBM's acoustic feature response is represented by each UBM component's statistics on emitting a feature vector ψ_t (of the t -th speech segment), namely the Gaussian posterior $\gamma_t(c)$; the zero order statistics N_c , accumulating component posteriors; and the first order statistics F_c , a component's expected value^a:

$$\gamma_t(c) = \frac{N(\psi_t | \mu_c, \Sigma_c)}{\sum_{c \in C} N(\psi_t | \mu_c, \Sigma_c)}, \quad N_c = \sum_{t \in T} \gamma_t(c), \quad F_c = \sum_{t \in T} \gamma_t(c) \psi_t. \quad (2.71)$$

Then, from the perspective of the enrolled reference term $\mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z}$, supervectors are centered by the UBM's supervector \mathbf{m}_{UBM} and the non-biometric supervector terms $\mathbf{U}\mathbf{x}$ —the origin of the supervector space is moved to $\mathbf{m}_{\text{UBM}} + \mathbf{U}\mathbf{x}$. As such, supervectors and first order statistics are centered as well; the centered supervector $\bar{\mathbf{m}}$ and the centered first order moment \bar{F} of a sample are denoted by (sparing the component-wise notation for the sake of easier tractability):

$$\bar{\mathbf{m}} = \mathbf{m} - (\mathbf{m}_{\text{UBM}} + \mathbf{U}\mathbf{x}), \quad \bar{F} = F - N(\mathbf{m}_{\text{UBM}} + \mathbf{U}\mathbf{x}). \quad (2.72)$$

For LLR computations, the JFA idea is to marginalize the observation in order to infer a latent subspace, where acoustic feature representations correspond to the biometric recognition task, i.e., the *latent biometric subspace*, which for JFA is jointly estimated with non-biometric factors. Therefore, the posterior distribution of the latent factors is estimated, such that the likelihood of a sample \mathcal{X} is computed by integrating over the posterior distribution of the sample (but not subject) depending term \mathbf{x} [129]:

$$\log \Pr(\mathcal{X} | \mathbf{m}) = \int \Pr(\mathcal{X} | \mathbf{m}, \mathbf{x}) \mathcal{N}(\mathbf{x} | \mathbf{o}, \mathbf{I}) d\mathbf{x}, \quad (2.73)$$

where \mathbf{o} represents the zero vector, and \mathbf{I} the identity matrix. Latent non-biometric factors are assumed to be Gaussian distributed (as well as the biometric and residual factors).

LLRs $S_{JFA}(\bar{\mathbf{m}}, \mathbf{x}, \bar{\mathbf{m}}_{UBM}, \mathbf{x}_{UBM})$ are estimated by the likelihood of a sample given a centered reference supervector $\bar{\mathbf{m}}$ and a probe's non-biometric factors \mathbf{x} and by the likelihood of this sample given the UBM's centered supervector $\bar{\mathbf{m}}_{UBM}$ and the UBM's non-biometric factors \mathbf{x}_{UBM} :

$$S_{JFA}(\bar{\mathbf{m}}, \mathbf{x}, \bar{\mathbf{m}}_{UBM}, \mathbf{x}_{UBM}) = \log \frac{\Pr(\mathcal{X} | \bar{\mathbf{m}}, \mathbf{x})}{\Pr(\mathcal{X} | \bar{\mathbf{m}}_{UBM}, \mathbf{x}_{UBM})}. \quad (2.74)$$

As: $\bar{\mathbf{m}}_{UBM} = \mathbf{0}$, the LLR computation simplifies where the log-likelihood terms are further approximated by the first order Taylor series; a linearly computable LLRs results [129]. Biometric and residual factors $(\mathbf{y}_r, \mathbf{z}_r)$ are extracted from reference samples, whereas non-biometric factors, zero and first order statistics $\mathbf{x}_p, \mathbf{N}_p, \mathbf{F}_p$ are extracted from probes. In JFA (for details, see [127–129]), the LLR $S_{JFA}(\mathbf{y}_r, \mathbf{z}_r, \mathbf{x}_p, \mathbf{N}_p, \mathbf{F}_p)$ is approximated by employing^b the UBM's covariance Σ_{UBM} [129]:

$$S_{JFA}(\mathbf{y}_r, \mathbf{z}_r, \mathbf{x}_p, \mathbf{N}_p, \mathbf{F}_p) \approx (\mathbf{V} \mathbf{y}_r + \mathbf{D} \mathbf{z}_r)^T \Sigma_{UBM}^{-1} (\mathbf{F}_p - \mathbf{N}_p (\mathbf{m}_{UBM} + \mathbf{U} \mathbf{x}_p)). \quad (2.75)$$

^a Second order statistics, accumulating $\psi_t \psi_t^T$ terms, cancel out during LLR computation and are thus spared in this brief overview [129].

^b Think of it as: $(\mathbf{V} \mathbf{y}_r + \mathbf{D} \mathbf{z}_r)^T \Sigma_{UBM}^{-1} (\mathbf{V} \mathbf{y}_p + \mathbf{D} \mathbf{z}_p)$ with probe biometric and residual probe factors $\mathbf{y}_p, \mathbf{z}_p$.

In JFA, voice samples are represented by a compound probabilistic embedding, comprising biometric, non-biometric, and residual embeddings. JFA is employed in text-dependent scenarios, where non-biometric factors resemble, e.g., different (known) passphrases uttered by a speaker or different capture sensors (microphones) used by a subject. The non-biometric JFA factors \mathbf{x} , however, were observed to convey speaker-discriminant information, such that the *i-vector* paradigm [130] emerged, see group (3) of Fig. 2.17.

Example: Probabilistic Embeddings, Acoustic i-vectors

In [130], a *total variability subspace* is proposed as a JFA special case. GMM supervectors \mathbf{m} are decomposed into the UBM supervector \mathbf{m}_{UBM} and the lower-dimensional representation—a latent variable referred to as *i-vector* \mathbf{i} :

$$\mathbf{m} = \mathbf{m}_{UBM} + \mathbf{T} \mathbf{i}, \quad (2.76)$$

where \mathbf{T} denotes the total variability matrix (a rectangular matrix of low rank), containing the eigenvectors with the largest eigenvalues of the total variability covariance matrix [130]. The total variability matrix can be thought of as a *dictionary*, mapping supervector elements to i-vector elements for each GMM component—the total variability matrix \mathbf{T} is constructed from component-wise entries \mathbf{T}_c with $c \in (1, \dots, C)$, giving a point estimate \mathbf{i} (an expectation) of the supervector \mathbf{m} by: $\langle \mathbf{m} \rangle = \mathbf{T} \langle \mathbf{i} \rangle$.

The total variability matrix is trained by the EM algorithm, in which i-vectors are modeled as latent Gaussian variables, more precisely, as Gaussian posteriors. The i-vector prior distribution $\Pi(\mathbf{i})$ with mean vector $\boldsymbol{\mu}$ and precision matrix \mathbf{P} is conventionally defined as [149] (note, precision is the inverted covariance):

$$\Pi(\mathbf{i}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{P}^{-1}). \quad (2.77)$$

The i-vector posterior distribution of $\Pi(\mathbf{i})$ with the i-vector point estimate $\langle \mathbf{i} \rangle$ and its uncertainty $\langle \mathbf{i} \mathbf{i}^T \rangle$ are given in terms of the component-wise pre-whitening^a zero and first order statistics N_c, \mathbf{F}_c by [149]:

$$\begin{aligned} \text{Cov}(\mathbf{i}, \mathbf{i}) &= \left(\mathbf{P} + \sum_{c \in C} N_c \mathbf{T}_c^T \mathbf{T}_c \right)^{-1}, \\ \langle \mathbf{i} \rangle &= \text{Cov}(\mathbf{i}, \mathbf{i}) \left(\mathbf{P} \boldsymbol{\mu} + \sum_{c \in C} \mathbf{T}_c^T \mathbf{F}_c \right), \\ \langle \mathbf{i} \mathbf{i}^T \rangle &= \text{Cov}(\mathbf{i}, \mathbf{i}) + \langle \mathbf{i} \rangle \langle \mathbf{i} \rangle^T. \end{aligned} \quad (2.78)$$

Conventionally, $\Pi(\mathbf{i})$ is assumed to be standard normal ($\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{P} = \mathbf{I}$), simplifying the i-vector extraction to:

$$\langle \mathbf{i} \rangle = \left(\mathbf{I} + \sum_{c \in C} N_c \mathbf{T}_c^T \mathbf{T}_c \right)^{-1} \left(\sum_{c \in C} \mathbf{T}_c^T \mathbf{F}_c \right). \quad (2.79)$$

^a After pre-whitening, the precision matrix associated with each mixture component can be taken to be the identity matrix and the mean vector to be zero [149].

Notably, i-vectors are point estimates associated with their depending extraction uncertainty ($\langle \mathbf{i} \mathbf{i}^T \rangle$), which is employable in feature

space analyses and in comparison algorithms (for the latter, see [53, 137]). Acoustic i-vectors are also employable to other tasks outside of speaker recognition, such as speech, language, and emotion recognition. For the purpose of this dissertation, acoustic i-vectors are used in order to derive *q-vectors*, which are proposed to inform score calibration about changing environmental conditions. For speaker recognition, however, decisions are carried out in the biometric (not the acoustic) subspace, thus acoustic i-vectors are projected into the biometric subspace.

Example: From Acoustic to Biometric i-vectors

State-of-the-art i-vector systems, such as in [150], conduct a variety of i-vector post processing in order to enhance the biometric discrimination performance. Linear discriminant analysis (LDA) transforms a feature space by maximizing the ratio of the class between variance to the class within variance, targeting better linear class separability. i-vectors are already in the biometric subspace hereof. However, further feature normalizations lead to better recognition performance.

Whitening (or its supervised variation WCCN) decorrelates and denoises i-vector elements, such that its subspace resembles a Euclidean space. Thus, conventional distance measures are employable, e.g., the *Euclidean distance* and the *cosine distance*. After these transform, the covariance matrix of the data is the *identity matrix*.

In [132], the length of i-vectors is observed to be proportional to the speech duration; more speech segments correspond to more sufficient statistics accumulated during i-vector extraction and thus to more sufficiently estimated i-vector distributions. For the purpose of better comparability of i-vectors from variable voice sample durations, [132] proposes to employ radial Gaussianization, particularly its Taylor series approximation *length-normalization*. Effectively, i-vectors are put to unit length and are thus projected onto a unit sphere around the feature space's origin—length-normalization in the Euclidean space is referred to as *l2-norm*.

If desired^a, i-vector posterior distributions can be projected throughout LDA, WCCN, and l2-norm [151].

^a The vast majority of i-vector systems is not propagating uncertainty to the comparator, saving computational efforts.

2.5.4.2 DNN Embeddings

Different deep learning schemes are studied within the speaker recognition community, whereof the x-vector embedding [134, 135] yielded the greatest (recent) breakthrough. In this dissertation, discriminative embeddings are briefly outlined as part of the literature survey on speaker recognition, however, they are not investigated thereafter as

the data interpretability is diluted during DNN processing. Among others, the preservation of interpretability is relevant to forensic scenarios.

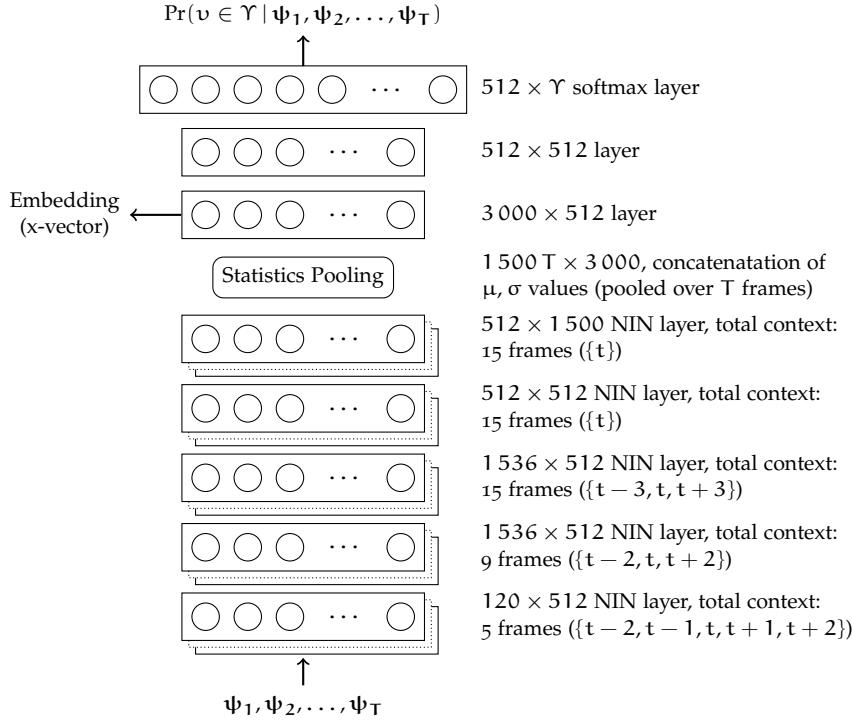


Figure 2.20: Architecture of end-to-end DNN extracting x-vector embeddings at the sixth layer before the nonlinearity, see [136]: layers operate on T acoustic features in a feed-forward DNN; a statistics pooling layer aggregates over the frame-level representations with additional layers before a softmax over Υ training data subjects. During x-vector extraction, the last two layers are omitted, yielding 4.4 billion parameter estimations.

Example: DNN Embeddings, x-vectors

End-to-end DNNs extract embeddings (referred to as x-vectors) [134–136] derived from acoustic features, namely 20 static MFCCs without derivatives. In [136], 24 filterbank responses—mean-normalized with sliding windows up to 3 s [145]—are derived from speech segments (after VAD). Fig. 2.20 illustrates the x-vector DNN architecture.

In each layer, DNNs employ linear functions with input/output vectors \mathbf{x} , \mathbf{y} , projection matrix \mathbf{A} , and bias \mathbf{b} :

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}. \quad (2.80)$$

Network layers are normalized by activation functions (e.g., the ReLU activation function), typically introducing a nonlinearity to the network. Parameters \mathbf{A} , \mathbf{b} are trained by solving Lagrangians, optimizing an objective function of the recognition task. Ideally,

outputs of higher layers (before nonlinearities) convey more discriminant features for the recognition task at hand, such as x-vector embeddings for speaker recognition.

The DNN employs five layers with a *network-in-network nonlinearity* (NIN), see [152]. Effectively, where conventional DNNs would have a rectified linear unit (ReLU) activation function, NINs propose three ReLUs with two so-called *micro neural network blocks* in between. Thereby, NIN layers assess different temporal contexts: the first NIN layer uses two frames before and after a current frame (a context of five frames with 120 inputs, yielding 512 outputs), the second NIN layer uses three outputs of the first layer (nine frames context with 1 536 inputs, yielding 512 outputs), and the third layer uses three outputs of the second layer (fifteen frames context with 1 536 inputs, yielding 512 outputs). After the fourth and fifth NIN layer, mean and standard deviation statistics μ, σ are aggregated over all frames, where the statistics pooling layer can, e.g., assign equal weights to frames or employ a self-attention mechanism [136, 153, 154]. Embeddings are extracted in the layer after the statistics pooling layer before its nonlinearity.

During training, additionally, a *feed-forward* layer and a softmax layer are employed. For each subject $v \in \Upsilon$ present in the training data $\{\psi_{1,l}, \dots, \psi_{T_l,l}\}_{l \in \mathcal{L}}$ with L voice samples, the softmax estimates the posterior probabilities, e.g., for the training identities $\Upsilon = \{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\} : \Pr(\mathcal{X} | \psi_1, \dots, \psi_T) = \frac{\exp(\mathbf{x}^T \mathbf{w}_\mathcal{X})}{\sum_{j \in \{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\}} \exp(\mathbf{x}^T \mathbf{w}_j)}$ with identity depending weights $\mathbf{w}_\mathcal{X}, \mathbf{w}_\mathcal{Y}, \mathbf{w}_\mathcal{Z}$. The training objective (the *loss function*) is the multiclass cross-entropy E :

$$E = - \sum_{l \in \mathcal{L}} \sum_{v \in \Upsilon} \theta_{l,v} \log \Pr(v | \psi_{1,l}, \dots, \psi_{T_l,l}), \quad (2.81)$$

where $\theta_{l,v}$ represents the ground-of-truth class label (1 if the subject of the sample l is v , and 0 otherwise) with a subject's posterior probability $\Pr(v | \psi_{1,l}, \dots, \psi_{T_l,l})$ given the evidence (the speech segments) $\{\psi_{1,l}, \dots, \psi_{T_l,l}\}$ having T_l speech segments. As there are many DNN designs and training objectives (e.g., see [155]), the x-vector recipe serves as a state-of-the-art prototype DNN.

2.5.5 Comparators for i-vector and x-vector Embeddings

In speaker recognition, the most prominent embedding comparators are the cosine distance similarity and subspace identity models, i.e., PLDA. Meta-embeddings [54, 99] are motivated by PLDA and x-vector embeddings, providing LLR comparisons to DNNs. For the sake of easier tractability, reference and probe embeddings are denoted by $\mathbf{x}_r, \mathbf{x}_p$ throughout this section.

Example: Cosine Comparison

In order to discriminatively compare i-vectors and x-vectors, the cosine between reference and probe features reveals their correlation.^a The cosine similarity score $S_{\cos}(\mathbf{x}_r, \mathbf{x}_p)$ is computed as:

$$S_{\cos}(\mathbf{x}_r, \mathbf{x}_p) = \frac{\mathbf{x}_r^T \mathbf{x}_p}{\|\mathbf{x}_r\|_2 \|\mathbf{x}_p\|_2}, \quad (2.82)$$

which for length-normalized embeddings simplifies to the inner product: $\mathbf{x}_r^T \mathbf{x}_p$, relating to the *Euclidean distance* $S_{\text{Euc}}(\mathbf{x}_r, \mathbf{x}_p)$ as: $S_{\text{Euc}}(\mathbf{x}_r, \mathbf{x}_p) = 2(1 - \mathbf{x}_r^T \mathbf{x}_p)$.

^a For training a discriminative embedding extractor, the cosine distance similarity between training samples is used. If its variation, the *angular margin softmax loss* [155, 156], is used instead, the DNN is trained in an elegant way to obtain well regularized loss function by forcing learned features to be discriminative on a hypersphere manifold [155] (further performance gains are promising).

The cosine comparator, however, neither examines the latent identity subspace, assuming extracted features are already discriminative for the recognition task at-hand, nor does it provide well-calibrated scores. These theoretical disadvantages manifest in higher decision costs, see [decision cost functions \(DCF\)](#). Whereas one might calibrate cosine scores to yield [LLRs](#), another might just compute LLRs directly.

State-of-the-art speaker recognition comparators belong to the [probabilistic linear discriminant analysis \(PLDA\)](#) family [132, 157, 158]. PLDA conducts an LLR scoring, comparing the probabilities of the propositions (a) reference and probe embeddings $\mathbf{x}_r, \mathbf{x}_p$ stemming from the same source or (b) stemming from different sources, see \mathcal{A}, \mathcal{B} in section 2.4.

Subject identities are modeled as latent variables in a linear subspace, assuming Gaussian distributions⁶⁹ for subject identities and noise [60]. In [60, chapter 18], the PLDA family is depicted. In this overview, emphasis is put on:

- the identity subspace model [60]
- the originally proposed PLDA [60]
- the simplified Gaussian PLDA [132, 158, 160] adapted for speaker recognition purposes
- the depending full subspace variant, i.e., the [two covariance model \(2Cov\)](#) [158, 160, 161].

⁶⁹ In [159], heavy-tailed priors (t-distributions) are proposed, accounting for pooled telephone and microphone speech.

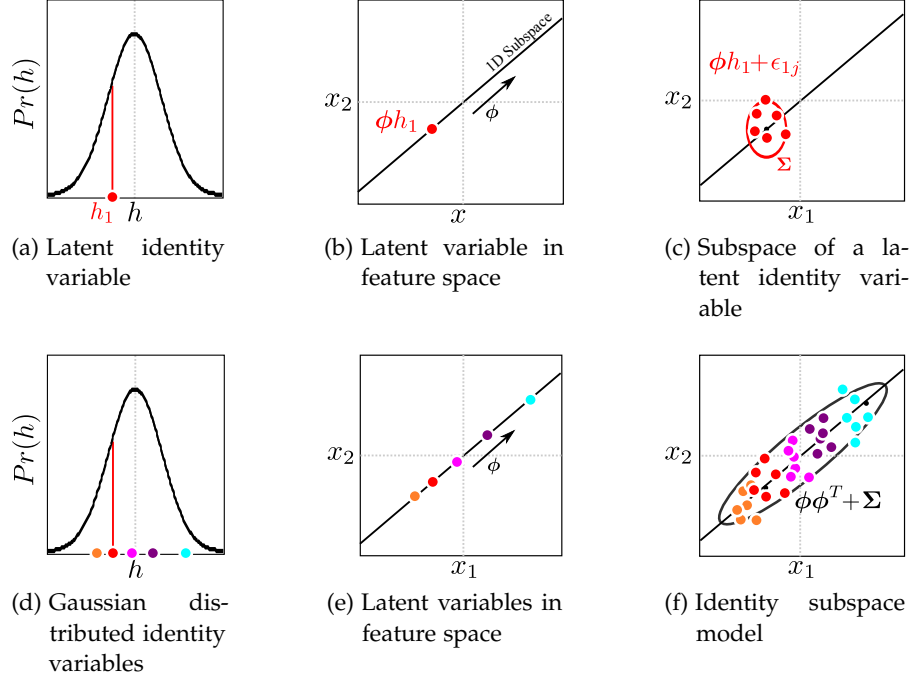


Figure 2.21: Identity subspace model, cf. [60].

Example: Identity Subspace Model Comparisons

The identity subspace model [60] extends canonical factor analysis by addressing within-subject variabilities. The generative model is depicted regarding \mathbf{x}_{ij} for I subjects with J_i samples per subject:

$$\begin{aligned}
 \mathbf{x}_{ij} &= \boldsymbol{\mu} + \boldsymbol{\Phi} \mathbf{h}_i + \boldsymbol{\epsilon}_{ij}, \\
 \Pr(\mathbf{x}_{ij} | \mathbf{h}_i) &= \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\Phi} \mathbf{h}_i, \boldsymbol{\Sigma}), \\
 \mathbf{h}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\
 \boldsymbol{\epsilon}_{ij} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \quad \text{with} \quad \text{diag}[\boldsymbol{\Sigma}], \\
 \mathbf{x}_{ij} &\sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Phi} \boldsymbol{\Phi}^T + \boldsymbol{\Sigma}\right), \tag{2.83}
 \end{aligned}$$

where $\boldsymbol{\Phi} \boldsymbol{\Phi}^T$ corresponds to the between-subject variance, $\boldsymbol{\Sigma}$ to the within-subject variance, $\mathbf{0}$ is the zero vector, and \mathbf{I} denotes the identity matrix. The model parameters $\boldsymbol{\Phi}, \boldsymbol{\Sigma}$ are iteratively trained by the EM algorithm.

Fig. 2.21 illustrates the identity subspace model for a 2D feature space: different identities are assumed to be standard Gaussian distributed, (a) and (d), which spans an 1D subspace in the 2D feature space. In the 1D subspace, different identities are distributed depending on the between-subject variance, (b) and (e). Thereby, the samples of one identity vary depending on the within-subject variance, which may also resemble a residual term to the model, (c) and (f). Assuming the same between and within class variances for all subjects, i.e., a model generalizing over all identities, the latent

1D identity subspace is expanded to the observed 2D feature space, yielding the identity subspace model.

The identity posterior distributions are estimated, similarly to the **i-vector** model, regarding the precision of the identity subspace estimation \mathbf{L}_i , providing first and second order moments $\langle \mathbf{h}_i \rangle, \langle \mathbf{h}_i \mathbf{h}_i^T \rangle$ [60]:

$$\begin{aligned}\mathbf{L}_i &= \mathbf{I} + \mathbf{J}_i \mathbf{\Phi}^T \mathbf{\Sigma}^{-1} \mathbf{\Phi}, \\ \langle \mathbf{h}_i \rangle &= \mathbf{L}^{-1} \mathbf{\Phi}^T \mathbf{\Sigma}^{-1} \sum_{j=1}^{J_i} (\mathbf{x}_{ij} - \boldsymbol{\mu}), \\ \Pr(\mathbf{h}_i | \mathbf{x}_{i;1}, \dots, \mathbf{x}_{i;J_i}) &= \mathcal{N}(\langle \mathbf{h}_i \rangle, \mathbf{L}^{-1}), \\ \langle \mathbf{h}_i \mathbf{h}_i^T \rangle &= \mathbf{L}^{-1} + \langle \mathbf{h}_i \rangle \langle \mathbf{h}_i \rangle^T.\end{aligned}\quad (2.84)$$

LLR scores $S_{\text{SIM}}(\mathbf{x}_r, \mathbf{x}_p)$ are computed in the feature domain by examining the distribution of identity variables in the latent subspace with: $\mathbf{\Sigma}_{\text{tot}} = \mathbf{\Phi} \mathbf{\Phi}^T + \mathbf{\Sigma}$ [157, 158]:

$$\begin{aligned}S_{\text{SIM}}(\mathbf{x}_r, \mathbf{x}_p) &= \log \mathcal{N} \left(\begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_p \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{\text{tot}} & \mathbf{\Phi} \mathbf{\Phi}^T \\ \mathbf{\Phi} \mathbf{\Phi}^T & \mathbf{\Sigma}_{\text{tot}} \end{bmatrix} \right) \\ &\quad - \log \mathcal{N} \left(\begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_p \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{\text{tot}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{\text{tot}} \end{bmatrix} \right).\end{aligned}\quad (2.85)$$

Accounting for further style variations (*channel* variability in speaker recognition), the original PLDA [60, 131] models these variations by additive Gaussian random variables. For PLDA with Gaussian assumptions, the simplifications are proposed in [132, 133].

Example: Original and Simplified PLDA Comparisons

The original PLDA assumes additional style (channel) influences, which are interpreted as additive Gaussian random variables \mathbf{s}_{ij} that (smoothly) result in an additive manner to the generative model:

$$\begin{aligned}\mathbf{x}_{ij} &= \boldsymbol{\mu} + \mathbf{\Phi} \mathbf{h}_i + \mathbf{\Psi} \mathbf{s}_{ij} + \boldsymbol{\epsilon}_{ij}, \\ \Pr(\mathbf{x}_{ij} | \mathbf{h}_i, \mathbf{s}_{ij}) &= \mathcal{N}(\boldsymbol{\mu} + \mathbf{\Phi} \mathbf{h}_i + \mathbf{\Psi} \mathbf{s}_{ij}, \mathbf{\Sigma}), \\ \mathbf{h}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ \mathbf{s}_{ij} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ \boldsymbol{\epsilon}_{ij} &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \quad \text{with} \quad \text{diag}[\mathbf{\Sigma}], \\ \mathbf{x}_{ij} &\sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Phi} \mathbf{\Phi}^T + \mathbf{\Sigma}).\end{aligned}\quad (2.86)$$

A compound generative model is established for LLR scoring, where reference and probe embeddings are simply stacked (*compound embedding*) [60]:

$$\begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Phi} & \boldsymbol{\Psi} & \mathbf{0} \\ \boldsymbol{\Phi} & \mathbf{0} & \boldsymbol{\Psi} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{r,p} \\ \mathbf{s}_r \\ \mathbf{s}_p \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_r \\ \boldsymbol{\epsilon}_p \end{bmatrix}. \quad (2.87)$$

A simplified variant is employed in speaker recognition [132, 159]: the generative model is defined in terms of a subject-specific component $\mathbf{B}_i = \boldsymbol{\mu} + \boldsymbol{\Phi} \mathbf{h}_i$ and a sample-depending component $\mathbf{W}_{ij} = \boldsymbol{\Psi} \mathbf{s}_{ij} + \boldsymbol{\epsilon}_{ij}$, where all latent variables $\mathbf{h}, \mathbf{s}, \boldsymbol{\epsilon}$ are assumed to be statistically independent. The generative model of the original PLDA can be re-formulated w.r.t. $\mathbf{B}_i, \mathbf{W}_{ij}$:

$$\begin{aligned} \mathbf{x}_{ij} &= \mathbf{B}_i + \mathbf{W}_{ij}, \\ \mathbf{B}_i &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Phi} \boldsymbol{\Phi}^T), \quad \mathbf{W}_{ij} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi} \boldsymbol{\Psi}^T + \boldsymbol{\Sigma}) \\ &\text{with } \text{full}[\boldsymbol{\Psi} \boldsymbol{\Psi}^T] \text{ and } \text{diag}[\boldsymbol{\Sigma}], \end{aligned} \quad (2.88)$$

such that the model is simplified to [132]:

$$\begin{aligned} \mathbf{x}_{ij} &= \boldsymbol{\mu} + \boldsymbol{\Phi} \mathbf{h}_i + \boldsymbol{\epsilon}_{ij}, \\ \mathbf{h}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ \boldsymbol{\epsilon}_{ij} &\sim \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}) \text{ with } \text{full}[\tilde{\boldsymbol{\Sigma}}], \\ \tilde{\boldsymbol{\Sigma}} &= \boldsymbol{\Psi} \boldsymbol{\Psi}^T + \boldsymbol{\Sigma}. \end{aligned} \quad (2.89)$$

LLR scores S_{PLDA} are computed alike to Eq. (2.87) in the feature domain by examining the distribution of identity variables in the latent subspace (with the identity covariance $\boldsymbol{\Sigma}_{\text{within}} = \boldsymbol{\Phi} \boldsymbol{\Phi}^T$ and the total covariance $\boldsymbol{\Sigma}_{\text{total}} = \boldsymbol{\Phi} \boldsymbol{\Phi}^T + \tilde{\boldsymbol{\Sigma}}$) [132]:

$$\begin{aligned} S_{\text{PLDA}}(\mathbf{x}_r, \mathbf{x}_p) &= \log \mathcal{N} \left(\begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_p \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\text{total}} & \boldsymbol{\Sigma}_{\text{within}} \\ \boldsymbol{\Sigma}_{\text{within}} & \boldsymbol{\Sigma}_{\text{total}} \end{bmatrix} \right) \\ &\quad - \log \mathcal{N} \left(\begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_p \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\text{total}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\text{total}} \end{bmatrix} \right), \end{aligned} \quad (2.90)$$

and by assuming $\boldsymbol{\mu} = \mathbf{0}$:

$$\begin{aligned} S_{\text{PLDA}}(\mathbf{x}_r, \mathbf{x}_p) &= \mathbf{x}_r^T \mathbf{Q} \mathbf{x}_r + \mathbf{x}_p^T \mathbf{Q} \mathbf{x}_p + 2 \mathbf{x}_r^T \mathbf{P} \mathbf{x}_p + \text{const} \\ &\text{with } \mathbf{Q} = \boldsymbol{\Sigma}_{\text{total}}^{-1} - \left(\boldsymbol{\Sigma}_{\text{total}} - \boldsymbol{\Sigma}_{\text{within}} \boldsymbol{\Sigma}_{\text{total}}^{-1} \boldsymbol{\Sigma}_{\text{within}} \right)^{-1}, \\ &\quad \mathbf{P} = \boldsymbol{\Sigma}_{\text{total}}^{-1} \boldsymbol{\Sigma}_{\text{within}} \left(\boldsymbol{\Sigma}_{\text{total}} - \boldsymbol{\Sigma}_{\text{within}} \boldsymbol{\Sigma}_{\text{total}}^{-1} \boldsymbol{\Sigma}_{\text{within}} \right)^{-1}. \end{aligned} \quad (2.91)$$

By centering the embeddings, a zero mean can be assumed, i.e., $\boldsymbol{\mu} = \mathbf{0}$. The closed-form solution is derived via Eq. (2.66) in terms

of the inner product of the natural parameters of each distribution and the sufficient statistics of the compound embeddings. Therein, the constant term summarizes the probability density functions' normalization terms depending on $\Sigma_{\text{within}}, \Sigma_{\text{total}}$.

For the purpose of faster score computations, a further simplification of PLDA is proposed and related to other generative pairwise models in speaker recognition, as depicted in [53, 133, 137, 158], namely the **two covariance model (2Cov)**, which can be represented in terms of a pairwise SVM, avoiding the *major weakness of one-versus-all SVM training* [158]. Using 2Cov pairwise SVM, multiple comparisons can be conducted at once across subjects, whereas one-versus-all SVMs need to be trained per subject.

Example: Two-Covariance (2Cov) Comparisons

The 2Cov model represents the full-subspace PLDA [161], a simple linear Gaussian generative model is adopted with between- and within-subject covariance matrices \mathcal{B}, \mathcal{W} which have the same dimensionality as the feature vector \mathbf{x}_{ij} [160]:

$$\begin{aligned} \mathbf{x}_{ij} &= \boldsymbol{\mu} + \mathbf{h}_i + \boldsymbol{\epsilon}_t, \\ \mathbf{h}_i &\sim \mathcal{N}(\mathbf{h}_i | \boldsymbol{\mu}, \mathcal{B}) \quad \text{with} \quad \text{full}[\mathcal{B}], \\ \mathbf{x}_{ij} | \mathbf{h}_i &= \mathcal{N}(\mathbf{x}_{ij} | \mathbf{h}_i, \mathcal{W}) \quad \text{with} \quad \text{full}[\mathcal{W}], \\ \mathcal{B} &= \sum_{i=1}^I \frac{J_i}{N} (\mathbf{h}_i - \boldsymbol{\mu}) (\mathbf{h}_i - \boldsymbol{\mu})^T, \\ \mathcal{W} &= \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{J_i} (\mathbf{x}_{ij} - \mathbf{h}_i) (\mathbf{x}_{ij} - \mathbf{h}_i)^T \end{aligned} \quad (2.92)$$

with the total number of samples $N = \sum_{i=1}^I J_i$.

The closed-form solution to the 2Cov score $S_{2\text{Cov}}(\mathbf{x}_r, \mathbf{x}_p)$ computation is denoted w.r.t. within and between precision $\mathcal{W} = \mathcal{W}^{-1}, \mathcal{B} = \mathcal{B}^{-1}$ with mean $\boldsymbol{\mu}$ [133]:

$$\begin{aligned} S_{2\text{Cov}}(\mathbf{x}_r, \mathbf{x}_p) &= \mathbf{x}_r^T \boldsymbol{\Lambda} \mathbf{x}_p + \mathbf{x}_p^T \boldsymbol{\Lambda} \mathbf{x}_r + \mathbf{x}_r^T \boldsymbol{\Gamma} \mathbf{x}_r + \mathbf{x}_p^T \boldsymbol{\Gamma} \mathbf{x}_p \\ &\quad + \mathbf{c}^T (\mathbf{x}_r + \mathbf{x}_p) + k \\ \text{with } \boldsymbol{\Lambda} &= \frac{1}{2} \mathcal{W}^T \tilde{\boldsymbol{\Lambda}} \mathcal{W}, \quad \boldsymbol{\Gamma} = \frac{1}{2} \mathcal{W}^T (\tilde{\boldsymbol{\Lambda}} - \tilde{\boldsymbol{\Gamma}}) \mathcal{W}, \\ \tilde{\boldsymbol{\Lambda}} &= (\mathcal{B} + 2\mathcal{W})^{-1}, \quad \tilde{\boldsymbol{\Gamma}} = (\mathcal{B} + \mathcal{W})^{-1}, \\ \mathbf{c} &= \mathcal{W}^T (\tilde{\boldsymbol{\Lambda}} - \tilde{\boldsymbol{\Gamma}}) \mathcal{B} \boldsymbol{\mu}, \quad k = \tilde{k} + \frac{1}{2} \left((\mathcal{B} \boldsymbol{\mu})^T (\tilde{\boldsymbol{\Lambda}} - 2\tilde{\boldsymbol{\Gamma}}) \mathcal{B} \boldsymbol{\mu} \right), \\ \tilde{k} &= 2 \log |\tilde{\boldsymbol{\Gamma}}| - \log |\tilde{\boldsymbol{\Lambda}}| - \log |\mathcal{B}| + \boldsymbol{\mu}^T \mathcal{B} \boldsymbol{\mu}. \end{aligned} \quad (2.93)$$

By using the Frobenius inner product^a [133], Eq. (2.93) is stated as:

$$\begin{aligned} S_{2\text{Cov}}(\mathbf{x}_r, \mathbf{x}_p) &= \langle \boldsymbol{\Lambda}, \mathbf{x}_r \mathbf{x}_p^T + \mathbf{x}_p \mathbf{x}_r^T \rangle + \langle \boldsymbol{\Gamma}, \mathbf{x}_r \mathbf{x}_r^T + \mathbf{x}_p \mathbf{x}_p^T \rangle \\ &\quad + \mathbf{c}^T (\mathbf{x}_r + \mathbf{x}_p) + k \end{aligned}$$

$$\begin{aligned}
&= \mathbf{w}_\Lambda^T \varphi_\Lambda(\mathbf{x}_r, \mathbf{x}_p) + \mathbf{w}_\Gamma^T \varphi_\Gamma(\mathbf{x}_r, \mathbf{x}_p) \\
&\quad + \mathbf{w}_c^T \varphi_c(\mathbf{x}_r, \mathbf{x}_p) + \mathbf{w}_k^T \varphi_k(\mathbf{x}_r, \mathbf{x}_p) \\
&= \mathbf{w}^T \varphi(\mathbf{x}_r, \mathbf{x}_p)
\end{aligned}$$

$$\text{with } \varphi(\mathbf{x}_r, \mathbf{x}_p) = \begin{bmatrix} \text{vec}(\mathbf{x}_r \mathbf{x}_p^T + \mathbf{x}_p \mathbf{x}_r^T) \\ \text{vec}(\mathbf{x}_r \mathbf{x}_r^T + \mathbf{x}_p \mathbf{x}_p^T) \\ \mathbf{x}_r + \mathbf{x}_p \\ 1 \end{bmatrix} = \begin{bmatrix} \varphi_\Lambda(\mathbf{x}_r, \mathbf{x}_p) \\ \varphi_\Gamma(\mathbf{x}_r, \mathbf{x}_p) \\ \varphi_c(\mathbf{x}_r, \mathbf{x}_p) \\ \varphi_k(\mathbf{x}_r, \mathbf{x}_p) \end{bmatrix},$$

$$\mathbf{w} = \begin{bmatrix} \text{vec}(\Lambda) \\ \text{vec}(\Gamma) \\ \mathbf{c} \\ k \end{bmatrix} = \begin{bmatrix} \mathbf{w}_\Lambda \\ \mathbf{w}_\Gamma \\ \mathbf{w}_c \\ \mathbf{w}_k \end{bmatrix}, \quad (2.94)$$

where the $S_{2Cov}(\mathbf{x}_r, \mathbf{x}_p) = \mathbf{w}^T \varphi(\mathbf{x}_r, \mathbf{x}_p)$ formulation can serve to outline a pairwise SVM.

^a The inner Frobenius product denotes $\mathbf{x}_r^T \mathbf{A} \mathbf{x}_p = \langle \mathbf{A}, \mathbf{x}_r \mathbf{x}_p^T \rangle = \text{vec}(\mathbf{A})^T \text{vec}(\mathbf{x}_r \mathbf{x}_p^T)$, where $\text{vec}(\cdot)$ is the operator stacking matrices into a vector, and $\langle \mathbf{A}, \mathbf{B} \rangle$ is the dot product between matrices, see [133].

Motivated by PLDA, *meta-embeddings* [54, 99] are proposed for LLR computations using DNNs, see group (5) of Fig. 2.17. Thereby, uncertainty is propagated in a principled manner, accounting for the embeddings extraction uncertainty.

Example: Meta-Embeddings

By propagating uncertainty, recognizers operating in changing quality settings are informed on the precision of the estimated feature. In meta-embeddings, the latent identity variable \mathbf{z} is employed to estimate LLRs in the latent subspace (not in the observation domain but similarly to PLDA). However, the purpose of meta-embeddings $f(\mathbf{z})$ is to *describe* a latent embedding \mathbf{z} , *not to define* it [99]: the latent variable \mathbf{z} is not definable by meta-embeddings, but geometric and algebraic structures are rather employed for the purpose of estimating LLRs, such that parameters outlining meta-embeddings are stored in vector form.

Gaussian meta-embeddings (GMEs) $f_{\text{GME}}(\mathbf{z})$ are represented by the *sufficient statistics* $\mathbf{T}(\mathbf{z})$, the D, D^2 -dimensional parameters \mathbf{a}, \mathbf{B} , and their *natural parameters* $\boldsymbol{\eta}(f_{\text{GME}}(\mathbf{z}))$:

$$\boldsymbol{\eta}(f_{\text{GME}}(\mathbf{z})) = \begin{bmatrix} \mathbf{a} = (\mathbf{I} + \mathbf{B}) \boldsymbol{\mu} \\ -\frac{1}{2} \mathbf{B} \end{bmatrix}, \quad \mathbf{T}(\mathbf{z}) = \begin{bmatrix} \mathbf{z} \\ \mathbf{z}^T \mathbf{z} \end{bmatrix}$$

$$\text{with } \mathbf{a} \in \mathbb{R}^D, \quad \mathbf{B} \in \mathbb{R}^{D^2}, \quad \boldsymbol{\mu} = (\mathbf{I} + \mathbf{B})^{-1} \mathbf{a}, \text{ s.t.:}$$

$$f_{\text{GME}}(\mathbf{z}) = \exp \left(\boldsymbol{\eta}(f_{\text{GME}}(\mathbf{z}))^T \mathbf{T}(\mathbf{z}) \right) = \exp \left(\mathbf{a}^T \mathbf{z} - \frac{1}{2} \mathbf{z}^T \mathbf{B} \mathbf{z} \right), \quad (2.95)$$

where \mathbf{B} is the positive, semi-definite precision matrix. The terms \mathbf{a}, \mathbf{B} are derived by a DNN (similarly to x-vectors). GMEs are practical for Bayesian inference, relying on (standard normal) priors $\pi(\mathbf{z})$, such that the log-expectation of a meta-embedding $\log \mathbb{E}(\mathbf{a}, \mathbf{B})$ is derived by integrating out \mathbf{z} (with $\langle \cdot \rangle$ denoting the expectation operator) [99]:

$$\begin{aligned}\pi(\mathbf{z}) &= \mathcal{N}(\mathbf{z} | \mathbf{0}, \mathbf{I}) = \frac{\exp\left(-\frac{1}{2}\mathbf{z}^T \mathbf{z}\right)}{\sqrt{(2\pi)^D}}, \\ \log \mathbb{E}(\mathbf{a}, \mathbf{B}) &= \log \langle f_{\text{GME}} \rangle_{\pi} \\ &= \log \int_{\mathbb{R}^d} f_{\text{GME}}(\mathbf{z}) \pi(\mathbf{z}) d\mathbf{z} \\ &= \frac{1}{2} \mathbf{a}^T (\mathbf{I} + \mathbf{B})^{-1} \mathbf{a} - \frac{1}{2} \log |\mathbf{I} + \mathbf{B}|. \quad (2.96)\end{aligned}$$

LLR scores $S_{\text{GME}}(\mathbf{a}_r, \mathbf{B}_r, \mathbf{a}_p, \mathbf{B}_p)$ with reference GME $\{\mathbf{a}_r, \mathbf{B}_r\}$ and probe GME $\{\mathbf{a}_p, \mathbf{B}_p\}$ are computed using the pooled reference–probe GME^a ($\mathbf{a}_{\text{pooled}} = \mathbf{a}_r + \mathbf{a}_p$, $\mathbf{B}_{\text{pooled}} = \mathbf{B}_r + \mathbf{B}_p$) [54, 99]:

$$S_{\text{GME}}(\mathbf{a}_r, \mathbf{B}_r, \mathbf{a}_p, \mathbf{B}_p) = \log \frac{\mathbb{E}(\mathbf{a}_r + \mathbf{a}_p, \mathbf{B}_r + \mathbf{B}_p)}{\mathbb{E}(\mathbf{a}_r, \mathbf{B}_r) \mathbb{E}(\mathbf{a}_p, \mathbf{B}_p)}. \quad (2.97)$$

^a When propagating uncertainty in a principled manner, LLR computations also include *matrix inversions* and *determinants* (of the GMEs' covariance terms).

2.5.6 Score Normalization and Calibration: Improving Decision Making

In speaker recognition, score normalization aims not at yielding scores in numerical intervals but at (a) adaptively augmenting the biometric information conveyed in a score by using additional data, and at (b) sustaining LLR score properties for single-algorithm and multi-algorithmic systems (for one comparison subsystem or when fusing multiple comparison subsystems). Theoretically, PLDA provides well-calibrated scores in terms of LLRs. Score normalization techniques, however, were shown to increase discrimination performance [162, 163].

Example: Cohort based Score Normalization Methods

By comparing the reference–probe comparison score S to the scores of additional comparisons carried out on so-called *cohort* data—additional data to the reference and the probe—normalized *z-scores* S' are estimated based on the cohort scores' average μ and on the cohort scores' standard deviation σ [162, 163]:

$$S' = \frac{S - \mu}{\sigma}. \quad (2.98)$$

Effectively, cohort scores describe an empirical score distribution, depending on the reference and the probe under comparison. The

score normalization puts different reference–probe comparisons on the standard normal scale. In large-scale systems, cohort sets comprise multiple thousands of samples. Depending on the employed comparator, Eq. (2.98) might solely consider references, probes, or both (in a symmetric variation).

The speaker recognition community refers to *z-norm* when the normalization scheme of Eq. (2.98) is applied to reference samples, yielding μ_r, σ_r normalization terms. The community refers to *t-norm* (*test* normalization) when this scheme is applied to probe samples, yielding μ_p, σ_p normalization terms. Notably, *z-norm* and *t-norm* tend to Gaussianize score distributions [89]. Usually, *z-norm* and *t-norm* are combined, either one on top of the other [162], e.g., referred to as *z/t-norm*, or in parallel [163], referred to as *symmetric normalization* (*s-norm*), when assigning equal weights to the *z*-score and *t*-score. Adaptive variations are considered to be more robust: distribution tails are omitted, enforcing cohort score distributions to be rather Gaussian [162, 163], e.g., by selecting the top-250 cohort scores. The *as-norm* (*adaptive s-norm*) score S_{AS} is computed by [163]:

$$S_{AS} = \frac{1}{2} \left(\frac{S - \mu_r}{\sigma_r} + \frac{S - \mu_p}{\sigma_p} \right). \quad (2.99)$$

First and second moment statistics $\mu_r, \sigma_r, \mu_p, \sigma_p$ are derived based on the top- n $\mathfrak{R}, \mathfrak{P}$ scores from each of the two resulting score sets \mathfrak{R} representing reference–cohort and \mathfrak{P} representing cohort–probe scores. Thereby, the most competitive cohort scores are selected with adaptation towards reference and probe features, i.e., μ_r, σ_r are ideally estimated on probe-alike cohort data, whereas μ_p, σ_p are ideally estimated on reference-alike cohort data.

Depending on the comparator, different normalization techniques were shown to be useful when targeting low [false match rate \(FMR\)](#) regions: either normalization schemes are employed in series, e.g., *z/t-norm* for [GMM–UBM](#) systems [162] and *as-norm* for [PLDA](#) systems [130, 163, 164].

Usually, cohort normalized scores are not LLRs, which is why score calibration sustains LLR properties [24, 28, 115]. Notably, score calibration is also useful for cosine or other non-LLR comparison approaches. Among others, score calibration methods are parametric (here, logistic regression) or non-parametric (here, isotonic regression) [165].

Example: Score Calibration by Logistic Regression

Logistic regression (or *linear calibration*) employs bias and scaling terms w_0, w_1 , estimated to provide a robust (but non-ideal) calibration of a score S . An activation as the linear combination of weights and score(s) is optimized regarding the best fit to a predictive value representing the class labels [58, 60], particularly for

binary decisions, a value between 0 and 1 representing \mathcal{B}, \mathcal{A} class propositions. Utilizing the logistic *sigmoid* function $\sigma(x)$, activations and predictions are linked by the calibration function $a(S)$:

$$\begin{aligned} a(S) &= w_0 + w_1 S, \\ \sigma(a(S)) &= \frac{1}{1 + e^{-a(S)}}, \end{aligned} \quad (2.100)$$

which, after the optimization of w_0, w_1 in terms of maximum likelihood learning [58, 60], corresponds to the posterior probability (at maximum prior uncertainty, i.e., $\pi = 0.5$):

$$\Pr(\mathcal{A} | S) = \sigma(a(S)). \quad (2.101)$$

Therefrom, LLR calibrated scores S' are approximately obtained using the logit function^a:

$$S' \approx \text{logit}(\sigma(a(S))) = a(S) = w_0 + w_1 S. \quad (2.102)$$

^a The logistic sigmoid function is the inverse of the logit function, transforming log-odds to probabilities. Notably, neural networks designed with regard to a probabilistic generalization of embeddings [99] can be assumed well-calibrated. By contrast, logistic calibration [79, 165] resembles a single perceptron network.

As presented in [79] and further advanced in [32, 38, 166, 167], additional terms can be added in order to adaptively characterize a comparison and to increase discrimination and calibration performance. Regarding this dissertation, the extension of logistic regression directly motivates the research objective [71, 72]. As score calibration employing logistic regression sustains a parsimonious degree-of-freedom, robustness can be assumed over a limited range of effective priors. However, on rather low or high LLRs, approximation errors tend to increase. Eventually, the degree of freedom is a trade-off between reaching ideal calibration and sustaining robustness.

Example: Score Calibration by Isotonic Regression (PAV-LLR)

In score calibration, comparison scores^a are transformed to LLRs. The [pool adjacent violators \(PAV\)](#)-LLR algorithm is a special case of isotonic regression, see section 2.4.4. PAV establishes score groups representing alike class \mathcal{A}, \mathcal{B} proportions based on the empirical observation of system score outputs. The PAV calibration function is defined by the score group bounds and by the group depending LLR values. For empirical scores (e.g., outside of technology and laboratory testing) situated between score groups, PAV-calibrated LLR values are estimated by linear interpolation. Score calibration is isomorph w.r.t. discrimination performance but improves calibration performance as LLR score properties are sustained.

^a Binary decision LLR scores can be alike to *similarity* or *dissimilarity* scores, depending on the definition of the propositions \mathcal{A}, \mathcal{B} . Conventionally, and in this dissertation, *similarity scores* are computed.

The PAV algorithm [28] provides optimal calibration for a given dataset by conducting an isotonic regression, see Fig. 2.12. The calibration, however, is based on score group mappings and interpolations which might be prone to nuisance artifacts. In contrast to linear calibrations—solely achieving good calibration on a limited range of operating points—, PAV is capable of calibrating well over wide ranges of operating points due to its nonlinearity [165], cf. Fig. B.1j. Thus, PAV-LLR calibration is referred to as *oracle/ideal score calibration* [24, 115].

2.6 SUMMARY AND CONCLUSION

This chapter introduced different fundamentals relevant to this dissertation's discourse, particularly: *the narrative on paradigms in decision making*; *biometric systems in standardization*; *a gap analysis between biometric standardization and speaker recognition communities*; *the Bayesian decision framework (BDF)* (basis to speaker recognition system design and evaluation); and *the state-of-the-art in speaker recognition*. These topics interrelate with this dissertation as follows:

The Narrative on Paradigms in Decision Making. The scope of *where* and *how* a system is employed outlines its performance measures. For the purpose of promoting one amongst many recognition systems to be employed in *one* application operating in fixed environmental conditions of *high quality*, error rate based figures of merit can serve as a proxy criterion to information theoretic performance measures. From an information theoretic perspective, recognition system outputs as comparison scores aid the prediction of classes. A prior belief in the class observation probabilities and cost belief in the impact of erroneously made decisions outline application-dependent figures of merit. For defined prior and cost beliefs, comparison scores ideally predict class proportions that equal prior and cost proportions. Having the uncertainty in prior beliefs fixed, however, the trade-offs in error rates—error rates resemble proportions of comparison scores—correspond to trade-offs in cost beliefs. When researching and developing recognition systems for multiple application purposes, e.g., access border control, banking and payment solutions, and forensics, application-independent figures of merit assess the benefit of employing the same recognition system for each purpose. (Alternatively to optimizing one system across decision trade-off requirements, recognition system vendors might recalibrate system outputs for each service provider; however, this practice is believed to be too intractable within this dissertation.) Thereby, in contrast to commercial scenarios but by using the same application-independent figures of merit in forensic scenarios, systems can be optimized to better contribute for informed decision making; in forensics, levels of evidence reporting (source and activation levels)—the score outputs of an automated

recognition system—are separated from the level of decision making (the offense level), although cost beliefs remain unspecified. To serve all these different demands, scores need to be **log-likelihood ratios (LLRs)**. In this dissertation, the scores of speaker recognition systems operating in unconstrained environments are recalibration to sustain LLR scores.

Biometric Systems in Standardization. The general design of a biometric system distinguishes between subsystems: data capture, signal processing, data storage, comparison, decision, and **presentation attack detection (PAD)**. This dissertation contributes to all but the data capture subsystem—in speech technology, the variability of capture devices (microphones) is well-studied to be diminished by the normalization of acoustic features, such as *cepstral mean and variance normalization* of **mel-frequency cepstral coefficients (MFCCs)**. On the comparison subsystem, a greedy learning architecture is proposed in annex A to retrieve biometric information when only limited training data of mobile device speech is available. On the PAD subsystem, a countermeasure on *unit-selection* attacks is proposed in section 7.1. On the data storage subsystem, a biometric information protection scheme is proposed in section 7.2 based on which the architecture of distributed biometric systems depends and computations in signal processing, comparison, and decision subsystems are reformulated.

On the signal processing subsystem, **quality vectors (q-vectors)** are proposed and investigated in chapter 6 to inform comparison and decision subsystems on changing environmental conditions—in particular, on the impact of environmental changes to a specific signal processing. By this design choice, the standardized *quality control* module is substituted: in standardization, solely biometric samples of high quality are considered for comparison, assuming (i) characteristics could be recaptured, i.e., subjects are present during verification, e.g., in access border control, and (ii) the performance of biometric comparators is sustained, when lower quality features are not processed. For (i), however, recapturing might not always be possible, such as in forensic scenarios, where subjects are solely available during enrolment; at crime scenes, the biometric probe is a forensic trace. For (ii), this is true when extracting biometric features as templates; in speaker recognition, however, biometric features are extracted as models—the uncertainty associated with the feature extraction is propagated throughout signal processing and can be used in comparison subsystems. As for templates, uncertainty cannot be propagated in a principled manner, quality control is necessary. When propagating uncertainty to the decision subsystem, however, good decision making can be sustained despite degrading signal quality; the capability of using q-vectors for this purpose is investigated in this dissertation. During this dissertation’s research on q-vectors, the 2014 state-of-the-art in speaker recognition is employed as the base-

line system and the acoustic and biometric feature extraction is fixed. Therefore, the impact of *unconstrained environments* to the acoustic feature extraction is investigated in chapter 5.

On the decision subsystem, the perspective of requirement specification, communication and validation based on prior and cost beliefs is communicated to the biometric standardization (and other machine learning communities) in chapter 4. A proposed taxonomy on performance visualizations interrelates targeted audiences, perspectives on performance and depending criteria types. The [binary decision error trade-off \(BET\)](#) plot is proposed: changes in magnitude of prior and cost beliefs linearly resemble on this error rate trade-off's canvas. Consequently, changes in LLR thresholds are revealed; the BET plot visualizes requirement trade-offs in the latent decision subspace—betting log-odds are revealed.

Gap Analysis. Gaps of the biometric standardization to the speaker recognition community are depicted, particularly on performance evaluation. Benefits are outlined of employing [strictly proper scoring rules](#), the [BDF](#), and evaluation methodology motivated from information theory. Effectively, these gaps correspond to moving from a Frequentist to a Bayesian perspective on performance assessment. Exemplarily, impacts to the Frequentist [rule of 3](#) and [rule of 30](#) are shown, when moving to a Bayesian perspective. As the BDF is the formalized way of denoting thresholds, these examples correspond to specifying threshold values before an evaluation, error rate trade-offs lead to associated thresholds; and after an evaluation, prior and cost requirements directly yield a threshold's value. Both perspectives consider score *discrimination* performance, the latter also includes score *calibration* performance: well-calibrated systems lead to a decision risk that approximates the discrimination performance well. By employing worse calibrated systems, higher decision risks are the consequence.

The Bayesian Decision Framework (BDF). This gap is bridged by the BDF. Thus, the BDF is explained in-depth from basis in total probability theorem and identity inference; over the derivation of decision risk performance, i.e., application-dependent figures of merit; relations between the [receiver operating characteristic \(ROC\)](#) plot, the [ROC's convex hull \(ROCCH\)](#) and ideal score calibration; performance visualizations in forensic evaluation and motivated from information theory; to the application-independent figures of merit C_{llr} and C_{llr}^{\min} . Thereby, C_{llr}^{\min} reports on discrimination performance, stating the lower bound to C_{llr} , which also reports performance losses from (insufficient) calibration. C_{llr} and C_{llr}^{\min} are the primary performance measures used in this dissertation, as both summarize different perspectives on performance, particularly, generalized cross-entropy and decision risks, independently of threshold assumptions.

State-of-the-Art Speaker Recognition. A brief overview on breakthroughs, i.e., conventional methods, and on the state-of-the-art is outlined. The technological development of the speaker recognition community is summarized from the 1980s to 2019. The state-of-the-art technology used as baseline in this dissertation is explained in more detail, particularly, on the *i-vector* feature extraction and comparison by *probabilistic linear discriminant analysis* (PLDA). From the speaker recognition perspective, *biometric templates* are point estimates of *high precision*, whereas *biometric models* accommodate data uncertainty in the manner of varying signal contents and qualities—models analytically characterize the *rather uncertain information* (not on observed but) on inferred data. After reference–probe comparisons, speaker recognition systems normalize scores by employing additional cohort data (samples of known other speakers). Cohort samples are compared to either, references and probes, resulting in two score distributions, which are used to put the one score of reference–probe comparisons into context; cohort score normalization is adaptively to a comparison. If scores are not LLR scores (e.g., PLDA scores are LLRs but not anymore after cohort score normalization), score calibration is employed to transform (continuous) scores to LLRs, such that the BDF is persistently employable.

EXPERIMENTAL FRAMEWORK

This chapter introduces the experimental framework in terms of the evaluation methodology and the utilized datasets.⁷⁰ The datasets used encompass major datasets of the speaker recognition community, i.e., [speaker recognition evaluation \(SRE\)](#) datasets distributed by the [US National Institute of Standards and Technology \(NIST\)](#). This dissertation puts emphasis on a derived dataset from the 2012 NIST SRE, targeting different acoustic environments. The studies conducted for this dissertation on acoustic effects on voice segmentation as well as on limited mobile training data utilize the 2013 *MOBIO* SRE dataset originating from a European project, placing emphasis on speech in mobile environments. In this dissertation, protocols on privacy preservation for state-of-the-art voice comparators based on homomorphic encryption are proposed. For the purpose of providing an empirical validation of the implementation to the theoretical validity, the 2013–2014 NIST SRE database is used for the corresponding proof-of-concept study. The German speech data corpus (GSDC) of the Technische Universität Darmstadt and the S10 attack subset of the ASVspoof 2015 corpus are employed for investigating on [presentation attack detection \(PAD\)](#) countermeasures to the unit-selection attack.

3.1 EVALUATION METHODOLOGY

Experimental validations are carried out on relevant and well-established datasets. Evaluations employ well-established figures of merit of the speaker recognition and the biometrics standardization communities. The primary metrics on discrimination and calibration are C_{llr}^{\min} and C_{llr} , see section 2.4.7. As secondary metrics, error rate based metrics on discrimination are used, particularly the [equal error rate \(EER\)](#) and the [FNMR at a 1% FMR \(FMR₁₀₀\)](#), see section 2.4.4. In the following sections, the organization of data is introduced for system training and experimental validation, then performance criteria are depicted.

⁷⁰ Parts of the following dataset descriptions are derived from publications contributing to this dissertation [68–72, 75].

3.1.1 Organization of Data

For the purpose of preserving fair benchmarks, i.e., in order to avoid *data snooping*⁷¹ of the results presented (but also of future work building upon them), the data is organized in separated sets for different purposes:

- The studies on segmentation and on limited mobile training data utilize the MOBIO corpus [37, 168]. Thereby, the segmentation study synthetically adds noise in order to examine segmentation decision robustness in unconstrained environments. By contrast, the limited training data study assumes ideal conditions throughout but investigates on comparator training assumptions in the light of limited data availability during system development.
- The examined state-of-the-art speaker recognition system stems from the I4U consortium⁷² as of the 2012 NIST SRE [41, 169]. For training, it uses the [universal background model \(UBM\)](#), the [intermediate-sized vector \(i-vector\)](#) extractor, and the [probabilistic linear discriminant analysis \(PLDA\)](#) comparator. The I4U filelist on system development is used, whereas i-vectors are provided by I4U collaborators.⁷³
- Experiments on score normalization and calibration are carried out on the I4U file lists on development and calibration validation (not the 2012 NIST SRE evaluation dataset).
- The PAD study employs the ASVspoof 2015 corpus [49] and the GSDC provided by the TU Darmstadt [172].
- The proof-of-concept experiment on data privacy is conducted on the 2013–2014 i-vector NIST SRE [36], a remake of the 2012 NIST SRE with i-vectors extracted from a conventional system by the MIT Lincoln Labs and distributed by the NIST.

⁷¹ Data snooping refers to the discrepancy between pre-planned inference and inference after looking at data. In data driven research communities, new data needs to be regularly inserted into these communities, such that statistical errors are avoided. When running multiple analyses (within a research community), the *one* significant result of many other results could also be due to an observation sampling. This dissertation explicitly targets forms of development datasets for providing insights, such that research following this dissertation is not prone to data snooping.

⁷² The I4U consortium is a collaboration of institutes across four continents consisting of university and company labs. In 2012, nine institutes worked on a NIST SRE submission [169]. The file lists are prepared for sites participating in the 2012 NIST SRE as part of the I4U consortium. For the 2016 NIST SRE, the consortium comprised sixteen institutes [170]. For the 2018 NIST SRE, the consortium comprised eleven institutes [171].

⁷³ At this point, the author would like to give Rahim Saeidi a special thanks for his support.

A fixed acoustic signal processing is considered, i.e., i-vectors represent the acoustic (not yet the latent biometric) feature space. Experimental validations are carried out on I4U i-vector file lists. Transform functions to the latent biometric subspace are trained in accordance with the literature pool [32, 166, 167, 173].

3.1.2 Performance Criteria

The primary figures of merit concern the goodness of **log-likelihood ratio (LLR)** scores: C_{llr} for discrimination and calibration; and C_{llr}^{\min} for discrimination. Calibration loss as the gap from C_{llr}^{\min} to C_{llr} is referred to as the miscalibration cost of LLR scores C_{llr}^{mc} [166]: $C_{llr}^{\text{mc}} = C_{llr} - C_{llr}^{\min}$. As an application-independent discrimination metric, C_{llr}^{\min} represents the generalized empirical cross-entropy of class \mathcal{A} and class \mathcal{B} LLRs w.r.t. Bayesian thresholds $\eta \in (-\infty, \infty)$ on the assumption of well-calibrated systems [26, 115]. However, when systems are not well-calibrated, C_{llr}^{mc} increases⁷⁴, such that C_{llr} reports on the combined discrimination and calibration loss of a binary decision system.

The secondary figures of merit concern the biometric performance in terms of error rates and in accordance to the ISO/IEC 19795-1 [25] standard on biometric testing and reporting, namely the EER and the FMR₁₀₀ criteria.

3.2 DATASETS AND PROTOCOLS

In the following sections, employed databases, protocols and the related baseline performance are introduced. Systems are based on the conventional i-vector/PLDA paradigm, which is grounded on generative Bayesian models throughout the acoustic and biometric signal processing.⁷⁵

⁷⁴ Systems of low C_{llr}^{\min} costs are discriminative regarding the **Bayesian decision framework (BDF)** and systems of low C_{llr}^{mc} costs are also well-calibrated: C_{llr} approximates C_{llr}^{\min} . These systems may not need additional information in order to yield better C_{llr} performance. Conventionally, C_{llr}^{mc} occurs on lacking reference, probe, or model training data, e.g., when targeting environmental domains considered neither during the training of acoustic or biometric feature extraction nor during the training of comparators and score normalizations.

⁷⁵ In contrast to *discriminative* models, being solely optimized on a classification task at-hand, *generative* models are also capable of emitting the modeled data. Likelihoods depend on the model fit in the light of data being presented to this model. As such, generative models provide an estimate on their expectation when observing the data and are capable of maximizing this expectation by updating their parameters.

3.2.1 Protocol of the 2013 MOBIO SRE

Aiming at mobile environments, experiments are carried out on the publicly available 2013 MOBIO SRE [37, 168], which provides a challenging and realistic testbed for current state-of-the-art speaker verification [37]. The speaker recognition subset of the MOBIO database [37, 168] is recorded on mobile phones and laptops, whereas in the 2013 MOBIO SRE [37] only data from mobile phones is used. A standard acoustic feature extraction in speaker recognition is used based on the RAS-TAMAT [174] and the *jfcookbook* [175] toolkits: 60-dimensional acoustic features which are based on 19 mel-frequency cepstral coefficients (MFCCs) with log-Energy as well as first and second order derivative coefficients on a standard hamming window. Feature warping [145] is applied using a 3 s sliding window. Acoustic i-vectors are extracted with 400 dimensions based on a 512-component UBM. Due to the limited data in the 2013 MOBIO SRE, gender-independent systems are trained utilizing a PLDA comparator [132]: i-vectors are projected into a 49-dimensional latent biometric subspace by linear discriminant analysis (LDA), then mean values are subtracted, followed by within class covariance normalization (WCCN) [148] projection and length normalization [132]. Reference and probe i-vectors are compared by a full-subspace PLDA [60, 132, 158], known as the *two covariance model (2Cov)* comparator [133].

Table 3.1: Partitioning of the MOBIO database, cf. [37]

Set	Female		Male	
	Subjects	Samples	Subjects	Samples
Background	13	2496	37	7104
Development set (references)	18	90	24	120
Development set (probes)	18	1890	24	2520

Tab. 3.1 depicts the amount of subjects and samples for the training and development sets containing 50 and 42 subjects in total. Experiments are solely conducted on the development set in order to prevent data snooping effects for other research on the MOBIO database, targeting the MOBIO evaluation set. The MOBIO dataset is solely utilized for analyzing the performance of voice sample segmentation by voice activity detection (VAD) algorithms regarding noise simulations on different signal-to-noise ratio (SNR) levels. Due to a rather small proportion of female speakers in the PLDA training data, results are solely reported w.r.t. data of male speakers. The baseline system yields a 0.407 C_{llr}^{\min} , a 11.9% EER, and a 46.7% FMR₁₀₀ on the development set. The baseline EER performance is in line with other reported single-comparator system performances of 9.6% and 11.3% that solely use the MOBIO background dataset for system training,

including VAD. More elaborate multi-algorithmic system fusions employ larger datasets for training and achieve EERs up to 5.0%.

On the MOBIO evaluation set, the primary metric is referred to as the *half total error rate (HTER)* of the *false match rate (FMR)* and *false non-match rate (FNMR)* measured at the development dataset's EER threshold [37, 168]. HTER as a figure of merit is a mix between the perspectives of biometrics standardization and the BDF, thus the interpretation of HTER is conflicting either way.⁷⁶ For the sake of consistency to related work on the MOBIO SRE, experimental evaluations tertiary report on the HTER, too.

3.2.2 I4U Protocol of the 2012 NIST SRE

Experiments using the I4U file lists [41] as of the 2012 NIST SRE are employed in studies on *duration-only*⁷⁷ impacts (sample completeness) as well as on mutual duration and SNR score normalization and calibration. Tab. 3.2 provides an overview on the NIST SREs from 1996 to 2012 regarding the employed datasets, targeted languages, the number of subjects, and the duration of probe samples. In the vast majority of reference samples, the duration is at least one minute. Over the past decades, the SREs targeted different challenges, to which new datasets are provided in each evaluation.

Early datasets encompass different phases of the Switchboard collection⁷⁸ [177]: Switchboard 1 (sw1), Switchboard 2 phases 1 to 3 (sw2p1, sw2p2, sw2p3), Switchboard 3 phases 1 and 2 (sw3p1, sw3p2), targeting different English dialects in the U.S. and cellular data in Switchboard 3. The Ahumada corpus [178] targets Spanish speech, whereas the FBI database [179] puts emphasis on different microphones used in police interviews. The MIXER corpora [180–

⁷⁶ In the BDF, the EER threshold is where the maximum *minimum DCF (minDCF)* is observed. The depending *decision cost function (DCF)* parameterizes weights of Type I and Type II error rates. When examining the performance on another dataset, the DCF has the same parameterization and the same *LLR* threshold. By contrast, HTER applies an equal weight of both error rates and as the threshold remains fixed with no inherent or consistent *LLR* interpretation. In biometric standardization, the concept of weighting error rates is deprecated in [93]. Moreover, the current revision process of ISO/IEC 19795-1 [25] aims at deprecating the EER as well, since the EER might be the easiest to understand but remains the least informative figure of merit, among alternatives.

⁷⁷ Before investigating on mutual duration and noise effects, this dissertation investigates on duration impacts. The term *duration-only* indicates that unnoisy but truncated speech data is used.

⁷⁸ As of 2019, the Switchboard corpora still remain among the most employed evaluation corpora for various recognition tasks for telephone speech data as they comprise about 2 400 dialog conversations with phonetic transcriptions. From the speech processing perspective, the vast data amount of natural speech data transmitted over limited telephone bandwidth is rather relevant (compared to data captured by the latest microphones), such that these corpora are still included as training data in NIST SREs, among others.

Table 3.2: Overview on NIST SRE datasets, cf. [21, 27, 33, 34, 176]. Languages are encoded as English (E), Spanish (S), Arabic (A), Chinese (C), Russian (R).

NIST SRE	Databases	Languages.	N ^o subjects	Probe duration
1996–1999	sw2p3	E	233	3–60 s
2000	sw2p1, sw2p2, Ahumada	E, S	804	5–60 s
2001	sw1, sw3p1, Ahumada	E, S	174	5–60 s
2002	sw2p2, sw2p3, sw3p2, FBI	E	330	5–60 s
2003	sw2p2, sw2p3, sw3p2	E	356	5–60 s
2004	MIXER	E, S, A, C, R	310	10 s, 30 s
2005	MIXER	E, S, A, C, R	526	10 s, 5 min
2006	MIXER	E, S, A, C, R	1 011	10 s, 5 min
2008	MIXER	E, S, A, C, R	1 328	10 s, 5 min, 8 min
2010	MIXER, Greybeard	E, S, A, C, R	5 460	10 s, 2 min, 5 min, 8 min
2012	MIXER, Remix	E, S, A, C, R	2 250	30 s, 100 s, 5 min

[184] put emphasis on multi-lingual speaker recognition (including non-natives), on transmission channels (landline, cellular and microphone) as well as on various noise conditions. The latter being especially targeted in the 2012 NIST SRE, where subjects spoke in noisy booths. Other parts of distributed speech data contain synthetically added noise (after post-processing) [35]. The Greybeard collection [185] puts emphasis on aging by recapturing subjects of prior collections (Switchboard 1, Switchboard 2, Mixer 1, Mixer 3) at later time periods. In the Remix corpus [186], sample donors participating in earlier Mixer datasets are encouraged to make telephone calls from noisy environments, e.g., with street noise, music or television broadcasting speech in the background.

The I4U file lists for the 2012 NIST SRE are derived for the acoustic and for the biometric subspace modeling as:

- For the purpose of training the acoustic feature extraction with a 2048-component UBM, from which 400-dimensional i-vector are extracted, a variety of speech corpora are used (by I4U and other consortia participating in NIST SREs): Switchboard cellular phase 1 and 2 [177], Fisher English [187] of the NIST SREs 2001 to 2003 as well as the MIXER data [180, 181] of the NIST SREs in 2004 to 2006 [188–190].

- MIXER data [182–184] of the NIST SREs 2006 to 2010 [33, 34, 190] are employed for training the latent biometric subspace transform and the PLDA comparator with dimension reduction by LDA to 200, WCCN and length-normalization, and for full-subspace comparison by 2Cov. Different subsets are assembled for training and validating score normalization and calibration.

Table 3.3: Overview on the I4U calibration development (dev) and validation (val) sets regarding the references and probes subsets (ref, prb), cf. [41].

	N° subjects				N° samples			
	dev		val		dev		val	
	ref	prb	ref	prb	ref	prb	ref	prb
Male	680	868	763	804	16 941	19 866	29 961	21 837
Female	1 039	1 243	1 115	1 102	24 693	25 980	43 119	28 548

Tab. 3.3 provides an overview on the I4U development and validation sets to the anticipated calibration task. The 2012 NIST SRE evaluation plan encourages the use of other reference samples for recognition purposes. This, however, is in conflict with ISO/IEC 19795-1:2006 [191] (reconfirmed in 2016 [25]), prohibiting to employ data of other subjects of the evaluation test crew as well as data of other evaluation probes when conducting comparisons on the evaluation set. Results are solely reported on the I4U validation set.

3.2.2.1 Duration Study

In the duration-only study of this dissertation, the evaluation protocol is inspired by previous studies on the effect of voice sample duration effects in [32, 173] based on the I4U file list [41]. The system architecture and settings of the i-vector extractor used are reported in [169]. The used database subset contains 551 female and 425 male subjects with each subject having at least 10 samples. The focus on subjects with at least 10 samples is distinct to this dissertation’s study. Subject-disjunct development and validation subsets are separated into female and male reference and probe datasets. While both reference sets only contain full i-vectors, the probe sets contain truncated i-vectors of the duration groups 5 s, 10 s, 20 s, 40 s, and full (> 40 s) of each full sample after duration truncation.⁷⁹ The duration groups are depicted in Tab. 3.5a. Baseline performances are depicted alongside the mutual duration and noise study in Fig. 3.1.

⁷⁹ In the I4U file lists with duration truncation, some samples with less than 60 s but more than 40 s of speech were truncated into 5 s, 10 s, and 20 s durations. The original segment was put into the full condition group, leaving no 40 s representation of that sample. In existing literature [32, 173] and studies on duration indicate little difference between 40 s and full condition groups. Thus, this database property is assumed to cause negligible changes in the results of this dissertation.

3.2.2.2 Mutual Duration and Noise Study

In the mutual duration and noise studies, experiments are conducted for five duration and five SNR conditions. SNR conditions stem from two noise sources: air conditioning (AC) and crowd⁸⁰ (CROWD) noise. By degenerating voice samples from the 2012 I4U file list [41], mutual quality and completeness degradation effects are examined on 55 conditions on a state-of-the-art system comprising i-vector features [36, 130] and PLDA comparison [60, 132, 158]. Thereby, duration is considered to represent sample completeness, as the estimation of sufficient statistics, see section 2.5, becomes more precise (less uncertain) with the observation of more (speech) data. While in many scenarios, reference samples can be captured under very good conditions, probe samples are affected by signal degradation, hence emphasis is put on condition-variable probe samples, while references are assumed to be ideal.

Table 3.4: Label scheme for mutual duration and noise conditions.

		Condition								
		1	2	3	4	5				
		Duration	5 s	10 s	20 s	40 s	full			
		Noise	clean							
(a) Duration only conditions										

Condition	6	7	8	9	10	11 ... 15	16 ... 30	31 ... 55		
Duration	5 s			10 s			20 s ... full	5 s ... full		
Noise	AC					CROWD				
SNR	0 dB	5 dB	10 dB	15 dB	20 dB	0 ... 20 dB	0 dB ... 20 dB	0 dB ... 20 dB		
(b) Mutual duration and noise conditions										

Condition dependent voice sample versions are created from long duration and clean samples of the I4U file list [41] by truncation into the aforementioned duration groups as in [32, 173] and by applying AC and CROWD noise using the *Filter And Noise Adding Tool* (FaNT), such that noise groups of 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, and clean (original SNR) are established. In total, 55 conditions are examined, see Tab. 3.4.

The baseline speaker recognition system is set up as follows: *stabilized weighted linear prediction* [192] is employed for robust spectrum estimation after enhancing the voice sample data using *maximum-likelihood short-time spectral amplitude* [193]. The rest of the acoustic signal processing is similar to previous work in [32, 169]. Raw i-vectors are drawn from samples after VAD. The VAD labels from clean conditions are then applied to corresponding noise versions. For depend-

⁸⁰ Likewise babble noise.

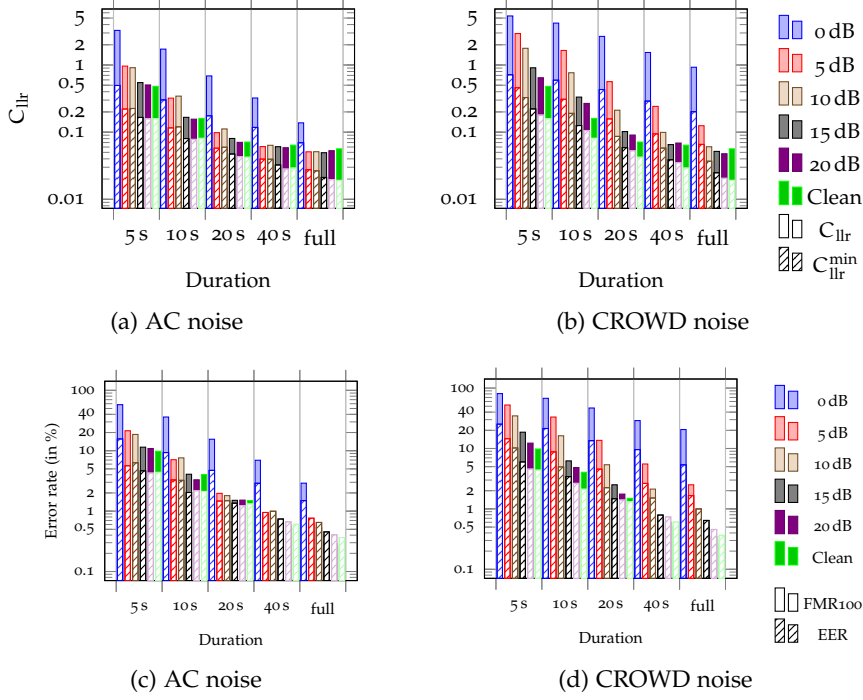


Figure 3.1: Baseline performance on I4U evaluation set affected by mutual signal degradation (without score normalization or calibration).

ing experiments, perfect VAD is assumed in order to exclude undesirable effects rising from VAD shortcomings in low-SNR levels which exceeds the scope of this dissertation. Full/clean probe samples of the I4U development and evaluation sets of all male subjects are synthetically modified condition-dependently. The conducted experiments assume similar performance effects to occur for female subjects as for male subjects.

All samples of the training set are modified condition-dependently, such that the latent biometric subspace feature processing produces well-calibrated scores over all conditions (but likely not for every single condition). The following processing is trained for acoustic i-vectors to yield discriminative feature representation in the latent biometric space: LDA reduces dimensions from 400 to 200, WCCN decorrelates feature elements and enforces unit variances. By length normalization [132], radial Gaussianization is approximated, projecting voice representation onto a unit sphere of i-vectors. For the sake of analyses tractability, experiments are solely reported regarding male speaker data. Biometric i-vectors are compared by PLDA with 200 speaker factors (full subspace; 2Cov comparator). PLDA is trained in a multi-condition pooled fashion as in [194].

Fig. 3.1 shows the performance of a state-of-the-art i-vector/PLDA baseline system with neither score normalization nor calibration. In general, CROWD noise causes higher performance deterioration than

AC noise. Longer duration and higher SNR levels lead to better performances, and in the vast majority of conditions, the clean/full condition outperforms other conditions in terms with $0.019 C_{llr}^{\min}$ and 0.4% EER (as expected). C_{llr}^{\min} performance vastly worsens due to high signal degradation on 0 dB SNR AC noise as well as on 0 dB and 5 dB SNR CROWD noise. By contrast, the performance of 20 dB SNR and clean conditions is rather similar on AC across duration conditions. By shifting the focus on duration effects, C_{llr}^{\min} and EER performance linearly depend on the log-duration (as observed by [32]) and mutual effects appear as a linear combination of log-duration and SNR impacts (SNR already is in log-compressed form).

3.2.3 2013–2014 NIST *i*-vector SRE Protocol

This dissertation’s proof-of-concept study on data privacy system architectures is carried out on the 2013–2014 NIST *i*-vector machine learning challenge [36, 195] phase III database (with labeled development data), where 600 dimensional *i*-vectors are supplied. The challenge dataset comprises a development subset of 36 572 *i*-vectors, 1 306 reference subjects with each having five enrolment *i*-vectors (references are average *i*-vectors), and 9 634 probe *i*-vectors. In total, 12 582 004 comparisons are carried out, jointly serving for two subsets on the progress of participants during the time of the challenge and the final evaluation performance. The challenge’s protocol [36, 195] resembles the 2012 SRE protocol [35], in which extracted acoustic *i*-vector features are distributed instead of audio data. The goal of the 2014 NIST SRE was to make research on speaker recognition available to a larger research community with diverse backgrounds in, e.g., machine learning, computer vision or other biometric modalities.

3.2.4 GSDC and ASVspoof 2015 Datasets

For the purpose of examining PAD of unit-selection attacks, the German Speech Data Corpus (GSDC) and ASVspoof 2015 datasets are employed. The ASVspoof 2015 corpus [49] only contains unit-selection [presentation attacks](#) in the evaluation set (as the *S10* attack); thus, a contrastive database is employed for development and optimization of countermeasures, namely the GSDC. As such, the ASVspoof data remains unseen, data snooping is prevented, and comparability with other countermeasures proposed at the spoofing challenge 2015 is preserved in the sense that the evaluation data is not exploited. However, by contrast, this dissertation’s study employs additional data to the data provided by the challenge; the ASVspoof 2015’s protocol would be violated for the sake of investigating on capabilities of detecting unit-selections if these [presentation attack instruments \(PAIs\)](#) were known. The training database on learning artefacts of unit-selec-

tion attacks is derived from the GSDC provided by the TU Darmstadt [172]. Tab. 3.5 depicts the dataset protocol.

Table 3.5: Database partitioning on GSDC (self partitioned) and ASVspoof 2015 (S10 eval-set).

Subset	Bona fide	Attack
GSDC development set	10 343	10 461
GSDC calibration set	3 745	4 484
GSDC validation set	400	100
ASVspoof S10 subset	9 404	18 398

3.3 SUMMARY

The used datasets are well-established within the speaker recognition community. For the purpose of examining effects to VAD and limited training data, the voice dataset of the 2013 MOBIO SRE is employed.

In most studies of this dissertation, experiments are carried out on a noise robust acoustic feature extraction originally based on file lists of the I4U consortium for the 2012 NIST SRE—data well-known to the speaker recognition community. Thereby, the employed acoustic and biometric feature extraction as well as the comparator are based on generative models which depict the conventional Bayesian paradigm in state-of-the-art speaker recognition throughout this dissertation. The parameterization of the baseline systems is in line with configurations reported in the research community and yields comparable performance.

Regarding evaluation criteria, results are primarily reported regarding discrimination power in terms of C_{llr}^{\min} and calibration power in terms of C_{llr} and C_{llr}^{mc} . Secondly, results are reported regarding Frequentist discrimination power in terms of [EER](#) and [FMR₁₀₀](#).

For research on PAD security, the performance of unit-selection attack countermeasures is assessed utilizing metrics proposed by the ISO/IEC standard 30107-3 [93], i.e., the [attack presentation classification error rate \(APCER\)](#) and the [bona fide presentation classification error rate \(BPCER\)](#). For research on privacy and data protection, i-vectors of the 2013–2014 NIST SRE are used that correspond to an i-vector re-distribution of the 2012 NIST SRE. This data is used for a proof-of-concept study on protocols proposed for preserving privacy.

ELABORATING ON THE PARADIGM SHIFT IN PERFORMANCE ASSESSMENT

This chapter addresses the paradigm shift in performance assessment from a Frequentist to a Bayesian perspective, see sections 2.3 and 2.4. The aim is to interrelate fundamental principles of the [Bayesian decision framework \(BDF\)](#) with a Frequentist's perspective on performance assessment. The following research question is targeted:

To which extent can the Bayesian paradigm on performance be interrelated with conventional error rate trade-off diagrams, and how can established performance visualizations be classified in the BDF?

The context of this chapter is outlined by well-established concepts within the speaker recognition, forensic sciences, and biometric standardization communities. For yielding contributions, interrelations are drawn to bridge gaps between these communities. Consequently, rather adequate methods in one research field might become less appropriate after the transfer to another field.⁸¹ As such, the contributions of this chapter on the BDF are more towards the general theory in machine learning than to a distinct applied research field (e.g., speaker recognition, forensics, and biometrics).

In order to make [log-likelihood ratios \(LLRs\)](#) digestible to the layman, especially in terms of visual performance assessment, error rate trade-off plots are examined regarding formal relationships towards [decision cost functions \(DCF\)](#)s, [empirical cross-entropy \(ECE\)](#), and C_{llr} metrics, see section 2.4. This chapter addresses the BDF and its implications to conventional performance visualization. Contributions to the theoretical framework are:

- **The formal notation and definition of an angular operating point.** It is well-known that LLR operating points resemble line segments in the [receiver operating characteristic \(ROC\)](#) space, the [ROC's convex hull \(ROCCH\)](#). As each segment is of constant slope, the turnover in LLR thresholds occurs in the angle between two ROCCH segments. From a Frequentist perspective, sampling error trade-offs correspond to moving thresholds;

⁸¹ Exemplarily, concepts from forensic science are motivated to make [likelihood ratios \(LRs\)](#) digestible to the layman in commercials, but in commercials, LRs are applicable in fully automated systems from evidence reporting to decision making. This is not the case in forensic evaluation, where the (semi-) automated evidence reporting (on source and activation level) is segregated from non-automated decision making (offense level). In contrast to commercials, the concept of LR thresholds is (at the moment) non-existent in forensic science.

from a Bayesian perspective, changing trade-off requirements corresponds to moving thresholds.

- **The visualization of verbal LLR scales in error rate trade-off plots.** Verbal scales for LLRs are introduced in the forensic science community in tabular form. To enhance evaluative transparency, these verbal scales are proposed to be visualized on the ROCCH.
- **The verbal scale of least-favorable decisions.** In forensic science, LLRs are communicated by *scales of strength of evidence*; priors cannot be assumed by forensic practitioners. There is a need to differentiate between the strength of evidence and the decision. In commercials, LLRs reflect not only the strength of evidence but LLR thresholds represent the least-favorable decision that can be made. To make LLRs as thresholds easier to digest in performance visualization, color-encoded bands of LLR threshold values are proposed in *detection error trade-off (DET)* plots. A host of infinite cross-application operating points is summarized into a few scales, which are easier to digest when deciding which system to favor, such that error trade-offs and Bayesian threshold values are interrelated. To forensic science, however, this visualization method corresponds to making LLRs easier to digest as scales of strength of evidence: since LLR scores and LLR thresholds are equal (at the least-favorable decision), either perspective holds. As *verbal scales of strength of evidence* already exist in tabular form proposed in the forensic science community, its visualization is a minor contribution. The visualization and contribution of *verbal scales of least-favorable decisions* is a more general task.⁸²
- **The introduction of a guideline on how to set and fine-tune decision policies formalized by (c, π) and thus LLR thresholds.** To make the BDF more applicable for practitioners, a communication scheme of formalized decision requirements is outlined. In brief, first a verbal band is selected, then an initial operating point is fine-tuned by relative adjustments of depending prior and cost beliefs.
- **The *binary decision error trade-off (BET)* plot.** The purpose of the BET plot is to visually interrelate changes in the magni-

⁸² In forensic science, the need to report on the strength of evidence results from the undefinability of priors and costs by forensic practitioners. Priors might be simulated, but for costs, only maximum uncertainty can be assumed. Nevertheless, in the BDF, LLR thresholds are outlined by prior and cost beliefs—even if one can use maximum uncertainty (value: 0.5, for binary decisions) to leave a belief unspecified—, and LLR scores need to be of equal value at least. The strength of evidence is summarized, mathematically, in the same manner in the ENFSI scale and the scale of conclusion as in the scales of least-favorable decisions. In theory, either represent the same aspect of the BDF. In practice, they differ in the way the BDF is applied.

tude of prior and cost beliefs with error rate trade-offs. Therefore, the axes of the y-axis inverted ROC plot are scaled by the *logit* transform; error rates are represented by their corresponding log-odds values instead of their conventional probabilistic representation. As such, error rate trade-offs are visually interrelated with LLR threshold trade-offs, which in turn correspond to the formal specification of application-depending decision policy parameterizations (c, π) , effectively revealing betting odds—and moreover, symmetry in trade-offs is preserved due to the treatment of their log-odds representation form.

- **The *normalized ECE (NECE)* plot.** ECE plots are normalized by visually accounting for the prior-depending default performance, such that visualized calibration losses are more directly comparable across different priors. To forensic evaluation, NECE plots visually remove prior dependencies. To commercial evaluation, NECE plots enable the simulation of information performance across priors independently of cost beliefs. For NECE plots, cost log-odds are effectively zeroed and thus neglected (only prior odds are considered) since nothing but the information gain is of interest, i.e., the information gain from employing a recognition system and its dependency on the (prior) proportion of the classes is to be recognized.
- **A self-contained (holistic) taxonomy on established and proposed BDF performance visualizations.** An overview is provided on the relationships and aspects of performance visualizations, particularly, for the performance reporting type (discrimination or discrimination and calibration), the performance reporting scope (error rates or information gains), and the targeted audience (analysis or reporting).

In the following, the ROCCH is favored over steppey ROC since it is consistent within the BDF, see section 2.4.4.

4.1 CONTRIBUTION: ANGULAR BAYES OPERATING POINTS

To visualize decision impacts, DCFs are revisited⁸³, see section 2.4. In this context, the parameterized Bayes risk as a DCF is computed for an empirical set of scores S given a specific operating point $\tilde{\pi}$ likewise η w.r.t. Type I and Type II error rates $p_1(\eta), p_2(\eta)$ as [28]:

$$\text{DCF}(S | \tilde{\pi}) = \tilde{\pi} p_2(\eta) + (1 - \tilde{\pi}) p_1(\eta). \quad (4.1)$$

Regarding DETs, the two addend terms can also be visually resembled in a two-dimensional Euclidean space, where each axis depicts an error rate, cf. [107].

⁸³ Parts of this section are based on a collaborative work with Daniel Ramos, Didier Meuwly, Jonas Lindh and Christoph Busch [66, 67].

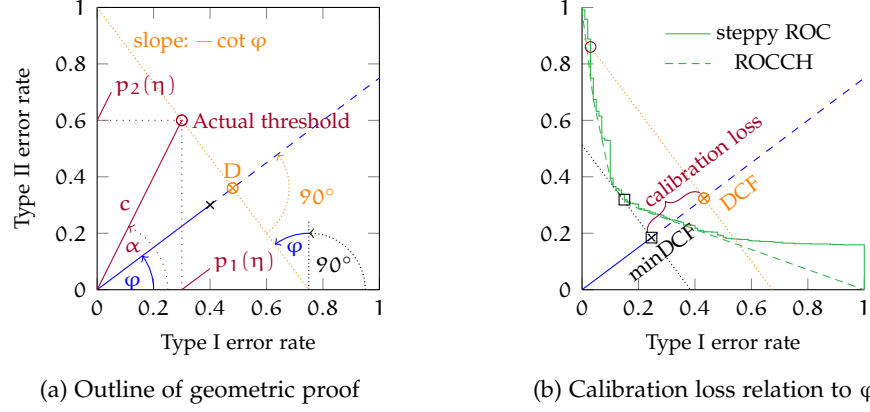


Figure 4.1: Operating points as linear combination in y-inverted ROC space: deriving DCF values by dropping perpendiculars onto the φ -depending line. Values of minDCF and DCF are derived by the distance to the origin of two lines having the slope $-\cot(\varphi)$ on an exemplary system. The first dotted line to the origin is the ROCCH tangent; the second dotted line intersects the φ -corresponding LLR threshold. The calibration loss resembles the distance between minDCF and DCF, cf. [107].

In other words, the Bayes operating point is a linear combination of the Type I and Type II error rates, and the ROCCH visualizes its **minimum DCF (minDCF)** for all $\tilde{\pi}$ associated operating points. Thus, before computing similarity scores, an operating point is denoted in the y-inverted ROC space by a line with $\tilde{\pi}$ depending slope. This dissertation proposes to denote the angular operating point φ by:

$$\tan \varphi = \frac{c}{1-c} \frac{\pi}{1-\pi} = \frac{\tilde{\pi}}{1-\tilde{\pi}} = \frac{\sin \varphi}{\cos \varphi} = e^{-\eta},$$

$$\text{DCF}(S | \varphi) = \sin(\varphi) p_2(\eta) + \cos(\varphi) p_1(\eta),$$

$$\text{note: } \varphi = \cot^{-1}(\tilde{\pi}^{-1} - 1), \quad \tilde{\pi} = (1 + \cot(\varphi))^{-1}. \quad (4.2)$$

Therefore, DCF slopes are depicted in terms of $\tan \varphi$ and $-\cot \varphi$. Regarding the $\tan \varphi$ slope, the DCF depicts a conventional linear function intersecting the origin, while the optimization task is subject to a minimal distance of its perpendicular towards the origin. The perpendicular is defined by the value of $-\cot \varphi$ and a y-axis offset. These $-\cot \varphi$ lines are tangents of the ROCCH for well-calibrated systems. Consequently, by sweeping over all φ values, the ROCCH resembles [79]. The notation of the angular operating point is implicitly depicted in [107]. Here it is explicitly denoted and defined.

Fig. 4.1 provides the outline of a geometric proof⁸⁴, cf. [107], a proof via ROCCH and minDCF relations is in [28]. Smaller φ values reflect more secure requirements, putting emphasis on ROCCH points of

⁸⁴ The proof directly follows from: $\frac{\overline{OD}}{c} = \cos(\alpha - \varphi) = \sin \varphi \sin \alpha + \cos \varphi \cos \alpha$, $p_1(\eta) = c \cos \alpha$, and $p_2(\eta) = c \sin \alpha$.

high Type II errors due to the $-\cot \varphi$ property and the convexity⁸⁵ of the ROCCH. On poorly calibrated systems, actual thresholds diverge from the threshold of minimum risk, i.e., the calibration loss as the difference between DCF and minDCF values is represented by the distance between the $-\cot \varphi$ lines of the actual threshold and of the ROCCH tangent.

Consequently, application-dependent operating points represent formalized decision policies and are specifiable in terms of the following notations (among further mixtures in between these notations, cf. section 2.4):

- unnormalized cost and prior parameters: $(c_A, c_B, \Pr(A), \Pr(B))$,
- normalized cost and prior parameters: (c_I, c_{II}, π) ,
- summarized cost and prior parameters: (c, π) ,
- effective priors (likewise, effective costs/policies): $(\tilde{\pi})$,
- angular operating points: (φ) ,
- LR thresholds: (τ) ,
- LLR thresholds: (η) .

All these expressions share the same meaning: an operating point formalized from a decision policy.

Considering the relationship between the ROCCH and PAV score calibration, angular operating points φ sample the PAV mapping function from the y-axis' perspective onto the 2D error rate trade-off diagram. The ROCCH results and in turn represents all Bayes risks by their discrimination power (all minDCF values, thus C_{llr}^{\min}). Since the relationship between the ROCCH, minDCF, and C_{llr}^{\min} is well-known, but the visual relationship to C_{llr} is spared so far in the literature, Fig. 4.2 exemplarily contributes the visualization of C_{llr}^{\min} and C_{llr} in the y-axis inverted ROC space. By sampling all points of the DCF on the related minDCF lines, the area visualizing C_{llr} is shaped.

4.2 VERBAL-SCALED DETECTION ERROR TRADEOFFS

This section⁸⁶ addresses the interrelation of performance visualization in terms of error rate trade-offs and LLRs. DET plots depict error rates independently of associated thresholds in quantile-quantile

⁸⁵ Pulling on one point of the ROCCH impacts all its segments as they monotonically preserve convexity, e.g., pulling on the [equal error rate \(EER\)](#) point would eventually result in a two-segment ROCCH—two straight lines to the left and right of the EER point. Other segments which would be in the concave set after pulling on a point of the ROCCH are removed and replaced by lines of the new convex set.

⁸⁶ Parts of this section are based on a collaborative work with Daniel Ramos, Didier Meuwly, Jonas Lindh and Christoph Busch [65–67].

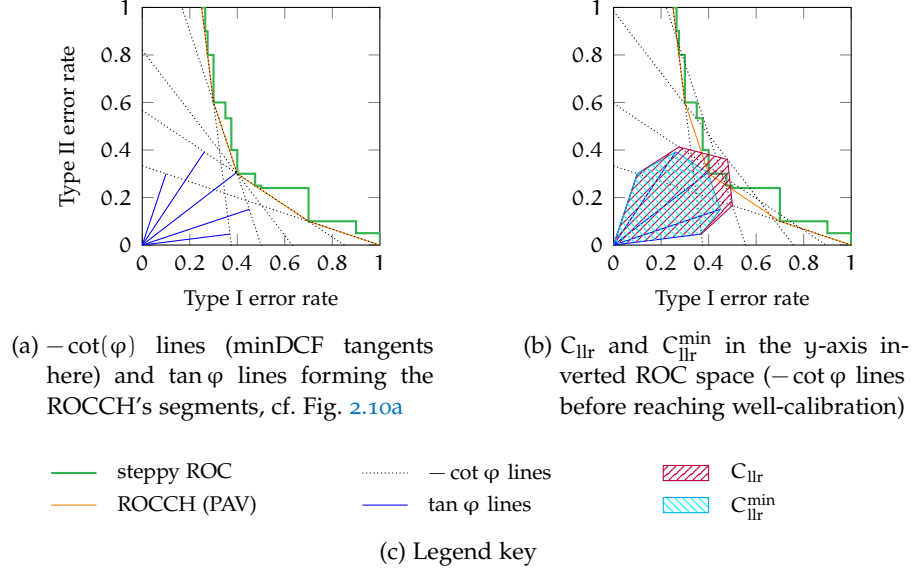


Figure 4.2: Contribution: visualizing C_{llr} in the y-axis inverted ROC space with (a) minDCF line and ROCCH tangent samplings, (b) visualizing C_{llr} areas in the y-axis inverted ROC space.

(Q-Q) plots, scaling y-axis inverted ROC plots. In contrast to the Frequentist approach of solely comparing error rates, the BDF further considers the impact of (wrong) decisions and the prior probabilities of each class, which form a BDF operating point (c, π) . In the forensic science community, magnitudes of LLRs are summarized by utilizing so-called *verbal scales* [23, 103, 196–200]. Verbal scales are introduced to differentiate between strength of evidence obtained from a holistic approach (verbal) and strength of evidence obtained from an empirical approach (numerical). LLR values quantify the strength of evidence, summarized in verbal scales by LLR magnitudes.

4.2.1 Making Likelihood Ratios Digestible: Verbal Scales

In forensic evaluative praxis, verbal scales are employed when LR and LLRs cannot be computed but one still needs to be able to provide a common scale for case work operating on (semi-) automated and fully manual examination and reporting of the weight of evidence. Notably, the consistency of human interpretation of LR values is shown to vary among (layman) examiners as well [201, 202], despite different effort levels in the training of (layman) examiners. As such, verbal scales aid a rather digestible and concise interpretation of LR values as verbal tags are defined in association to specific LR values, i.e., defining a LR range interpretation (by tags and values) for the purpose of human communication. In other words, verbal scales aid

the communication of system requirements and the issue to which extent a system's output supports either proposition.

Verbal LR scales map bands of LR values to verbal interpretation in terms of support for either prosecution or defendant propositions. Early verbal scales put emphasis on LRs $< 10^2$ [103, 200], until the forensic field considered LRs for DNA examination. In 2015, the European network of forensic science institutes (ENFSI) [23] recommended the verbal scale suggested by the association of forensic service providers (AFSP) [198] based on a ten digit system introduced by Turing. By contrast, Nordgaard et al. [199] proposed the *scale of conclusion*, which interpolates verbal bands concerning two fix points, namely $\text{LR} = 10^2$ and $\text{LR} = 10^6$. Thereby, the increase in the base 10 logarithm of two consecutive interval bands of the verbal bands is proportional. For the sake of easier tractability, these bands are abstractly labeled in terms of nine scales as $-4, \dots, \pm 0, \dots, +4$. In other words, the interpolation is conducted from the LLR perspective (in the log-domain, base e). This is more suitable than an interpolation within the LR-domain due to its linear symmetry regarding the depending scales of conclusion, where base 10 is considered for human emphasized assessment.

Table 4.1: Verbal scales for communicating LR values, cf. [200], scales depicted for LRs ≥ 1 . For LRs < 1 , bounds are symmetric as $\frac{1}{\text{LR}}$.

LR ≥ 1	verbal	LR ≥ 1	verbal	LR ≥ 1	verbal
≤ 3.16	barely worth	≤ 33	weak	$\leq 10^1$	limited
≤ 10	substantial	≤ 100	fair	$\leq 10^2$	moderate
≤ 31.6	strong	≤ 330	good	$\leq 10^3$	moderately strong
$\leq 10^2$	very strong	$\leq 10^3$	strong	$\leq 10^4$	strong
$> 10^2$	decisive	$> 10^3$	very strong	$> 10^4$	very strong
(a) Jeffreys'61 [103]		(b) Evett'91 [196]		(c) Evett'00 [197]	

LR ≥ 1	verbal	LR ≥ 1	verbal
$\leq 10^1$	weak/limited	≤ 5.625	(± 0) neither/nor
$\leq 10^2$	moderate	≤ 100	(+1) some extent
$\leq 10^3$	moderately strong	≤ 5625	(+2) support
$\leq 10^4$	strong	$\leq 10^6$	(+3) strong
$\leq 10^6$	very strong	$> 10^6$	(+4) extremely strong
$> 10^6$	extremely strong		
(d) ENFSI guideline [23, 198]		(e) Scale of conclusion [199] with non-approximated LR values	

Tab. 4.1 compares verbal scales on LRs ≥ 1 , which favor the proposition \mathcal{A} (prosecution). Verbal scales for LRs favoring the proposition \mathcal{B} (defense) with values < 1 are symmetric regarding $\frac{1}{\text{LR}}$. For the scope of this dissertation, the ENFSI scale is preferred for communication in forensic evaluative scenarios and the scale of conclusion for com-

munication in commercial scenarios. Either verbal scale is also illustrated in Fig. 4.3. The choice of a verbal scale depends on one's interest. Forensic reporting might prefer the ENFSI scale as the \log_{10} scale is easier to explain to laymen and provides more LR bands. By contrast, commercial reporting might favor the Nordgaard scale as the \log_e scale rather accounts for *natural* relations, and applications are denoted in a rather limited amount of bands that consider wider ranges of applications (which does not exclude its use to forensic science). For designing verbal scales, one might consider *Weber's law*⁸⁷—the more LR scores diverge from a (c, π) associated LR threshold, the wider LR bands might be; and the narrower if LRs are closer to the operating point (depending on the design and intention of involved decision policies). Notably, this dissertation solely outlines the applicability of *verbal scales* without explicitly intending to promote any particular scale design. Summarizing for the LLR domain, a verbal equivalent is used instead of an LLR, using a table for replacement. Thus, LLR values may be understood much more easily by end-users⁸⁸, conveying a degree of support rather than just a relative degree of similarity. This reporting scheme can favor end-users as *biometric system vendors, providers, operators, and owners*.

4.2.2 Contribution: Verbal Scales of Least-Favorable Decisions

Moving the perspective of verbal LLR scales to the domain of LLR thresholds (from forensic to commercial evaluation), this dissertation contributes to the verbal scale of *least-favorable decisions*, concerning an easier to understand summary of different decision policies. Fig. 4.4a shows a color code of the verbal scale [23] of LLR values, which is proposed to be used in error trade-off plots in order to interrelate LLR values at a given threshold in verbal bands. In forensic science (solely considering LLR values, not thresholds), such a figure could be used to show if a method is more suitable for investigation/intelligence or for evaluation. Associating verbal scales of LLR values to LLR thresholds, cf. Eq. (2.33), verbal scales are also put in context to BDF operating points (c, π) . LLR values that are greater than or equal to an LLR threshold η lead to the favoring of proposition \mathcal{A} at the minimal cost advantage.

⁸⁷ Effectively, the ratio of the *just noticeable difference* (JND) Δx perceived by a stimuli to its amount x is (roughly) constant, i.e., $\frac{\Delta x}{x} \approx \text{const.}$, inducing logarithmic JND effects regarding the stimuli amount. Exemplarily, let a difference in jail sentences be 3 months: from 3 to 6 months (*feeling* longer) compared to 20 years to 20 years and 3 months (*feeling* not that much more). The logarithmic stimuli might cause judges to speak less granular sentence terms and more round sentence terms (example also provided by YouTubers *Prof. Hannah Fry* and *Doctor of Letters Brady Haran*, see: <https://www.youtube.com/watch?v=hHG8io5qIU8>). When designing verbal scales for decision making, one might account for Weber's law when verbally tagging LR ranges (either on the base 10 or the natural logarithm).

⁸⁸ Even if it may be an illusion as words are less transparent than numbers.

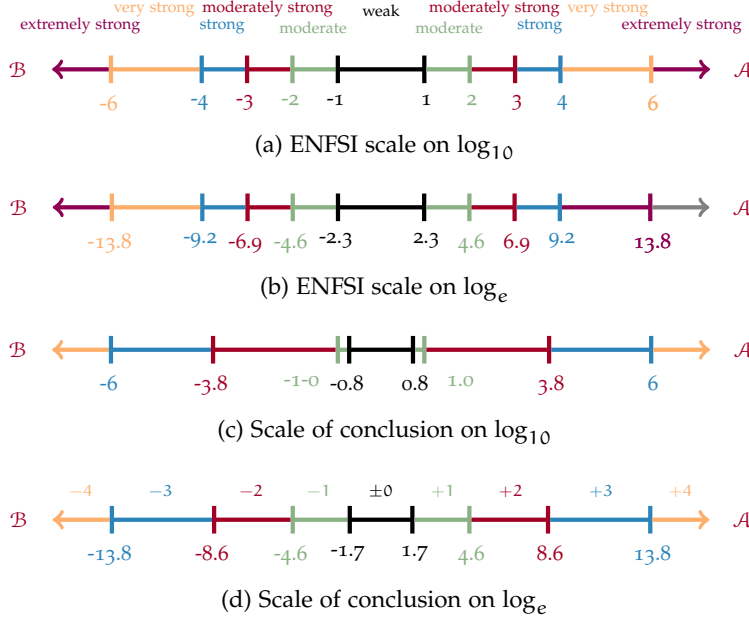


Figure 4.3: Comparison of verbal LR scales for base 10 and natural logarithm: verbal tags (above), associated log-value ranges (below).

4.2.3 Contribution: Verbal Scales in DET Plots

The bounds of verbal bands—of the strength of evidence and of least-favorable decisions—are depicted by using Eq. (4.2) in terms of the depending LR bounds⁸⁹ $\text{LR}_{-4,\dots,\pm 0,\dots,+4}$:

$$\tan \varphi = (\text{LR}_{-4,\dots,\pm 0,\dots,+4})^{-1}. \quad (4.3)$$

In order to visualize the cross-application discrimination performance in DET plots, verbal scales are utilized, e.g., the ENFSI scale and the scale of conclusion. Fig. 4.5 illustrates the verbal scales depending DCF slopes in the y-inverted ROC space, the y-inverted and log-compressed ROC space, and in the DET space. Since one minDCF point can lie on the φ depending line at most due to the ROCCH's convexity, levels of security and convenience are monotonically color-encoded on the ROCCH. Levels of decision policy requirements when aggregating applications by verbal scales are depicted.

The presented approach provides an assessment recipe towards cross-application decision risk for biometric researchers, hence DET plots are emphasized. This work proposes to depict aggregated levels of decision risk on ROCCHs w.r.t. verbal scales, assuming optimal calibration. Fig. 4.6 shows the verbal scale DET plot for three synthetic systems. The proposed visualization informs on both, the strength of evidence and the least-favorable decision—LLR scales are revealed

⁸⁹ This example refers to the scale of conclusion introduced in [199] since bounds are denoted from the LLR-domain, and the amount of bands is more limited to fewer categories, such that commercial decisions become easier to make.

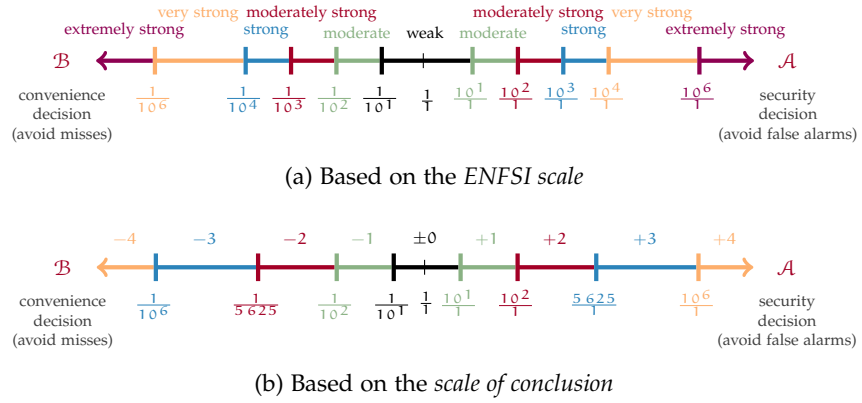


Figure 4.4: Contribution: verbal scale of *least-favorable decisions*: (a) translation of required $\log_{10}(\text{LR})$ values (annotated as the required posterior to prior odds ratio) into verbal support, based on [23], (b) translation of required \log_e values, based on [199]. Depending on the direction, LLRs will support decisions favoring \mathcal{A} or \mathcal{B} with a verbal strength indicated in the scale (alternating gray scales).

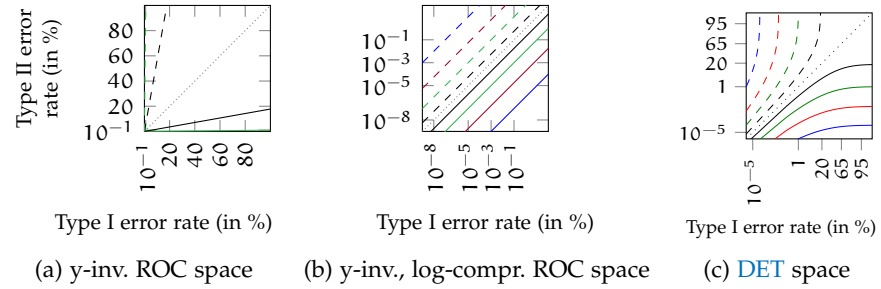


Figure 4.5: DCF slopes as $\tan \varphi$ lines in DET plots with φ operating points (exemplary on the scale of conclusion): solid lines indicate security levels (+1, +2, +3, +4), dashed lines indicate convenience levels (-1, -2, -3, -4). Blue, red, green and black lines indicate $\pm 4, \pm 3, \pm 2, \pm 1$ levels, respectively. The dotted line resembles as the center of the ± 0 level (DCF weights error rates equally).

in the latent decision subspace, where the values of LLR scores and LLR thresholds are the same. The nature of information, however, differs depending on an evaluation's purpose. In forensic evaluation, the strength of evidence provided by a system is of interest. In commercial evaluation, it is the validation of decision requirements.

Using the relationship between PAV and ROCCH, the proposed augmentation to DET plots increases transparency. Furthermore, the reflection of resulting PAV groups is motivated: ranges of supported verbal scales should be sustained, e.g., when binning within system processing or obfuscating scores. This is to be pursued to (a) yield ROCCHs and not make them collapse into a few supporting points and thus to (b) support a wider range of application requirements. Since the PAV conducts score binning, any pre-binning limits the

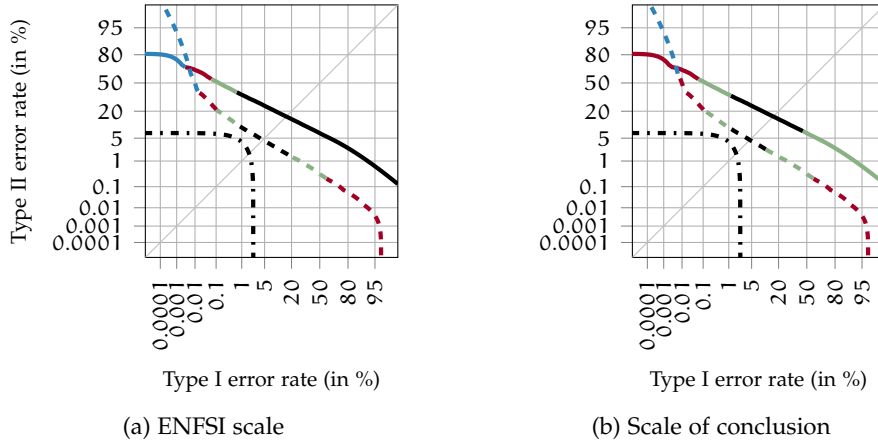


Figure 4.6: Contribution: verbal scales encoded DET plot with exemplary systems of Fig. 2.3a (solid), 2.3b (dashed), 2.3c (dash-dotted), and scales of least-favorable decisions based on (a) the ENFSI scale, (b) the scale of conclusion.

potential application range broadness of a system. The concept of depicting minDCF or DCF values in the DET space is not new, cf. application-independent evaluation methods [61], to which a novel scheme for depicting ranges of minDCFs is contributed here. They are aggregated by similarity in terms of verbal scales, suitable for comparing a few systems of interest. As a result of this work, error trade-offs are categorized into levels of effective security and convenience in terms of the decision domain. For forensic evaluation, the ENFSI verbal scale might be used instead of the scale of conclusion.⁹⁰ A publicly available reference implementation is provided.⁹¹

4.2.4 Contribution: Communicating Threshold Requirements

This section’s emphasis is placed on making the verbal scale encoded DET practical for commercial and forensic communication towards denoting BDF thresholds. Conceptually, communication parties might seek guidance as follows:

1. An application’s scale is denoted, such that a representative operating point can be assembled.
 Example. A security application of the verbal scale +2 is targeted. A lookup table, see Tab. 4.2, provides an initially associated threshold to each verbal band. For the scale +2, this thresholds is $\eta_{+2} = 8.03$.

⁹⁰ This debate does not lie within the scope of this dissertation, the ENFSI scale provides a higher variety of conclusion levels, whereas the scale of conclusion is easier to handle and visualizes a more natural summary of LLR magnitudes.

⁹¹ See: <https://codeocean.com/algorithm/154591c8-9d3f-47eb-b656-3aff245fd5c1/metadata>.

Table 4.2: Centers of $C_{llr}^{ratio}(\eta)$ gravity on the [199] scale of conclusion.

Verbal scale	-3	-2	-1	± 0	+1	+2	+3
Application	convenience				security		
η_{\min}	-13.82	-8.63	-4.61	-1.73	1.73	4.61	8.63
η_{center}	-13.18	-8.03	-4.07	0	4.07	8.03	13.18
η_{\max}	-8.63	-4.61	-1.73	1.73	4.61	8.63	13.82

2. A target prior⁹² π —one may think of it as an attack prior $(1 - \pi)$, i.e., the target prior for class \mathcal{B} —is denoted. Therefrom, the cost ratio $\frac{c_I}{c_{II}}$ is revealed by reformulating Eq. (2.33): $\frac{c_I}{c_{II}} = e^{\eta + \text{logit } \pi}$. Example. A friendly scenario has 99% a-priori target claims, such that $\pi = 0.99$. Therefrom, the cost ratio results as $\frac{c_I}{c_{II}} \approx 3 \times 10^5$. Note: when assuming $c_{II} = 1$ of a fixed but arbitrary cost unit, c_I is directly set by the same cost unit.
3. If the beliefs in π, c_I, c_{II} need adjustment, one can define mutability ranges of operative BDF thresholds η, η' in terms of an offset δ , which can be expressed as the logarithm of the odds ratio between the associated effective priors $\tilde{\pi}, \tilde{\pi}'$:

$$\begin{aligned}
\delta &= \eta' - \eta = \text{logit } \tilde{\pi} - \text{logit } \tilde{\pi}' \\
&= \log \frac{c_I'}{c_{II}'} - \text{logit } \pi' - \left(\log \frac{c_I}{c_{II}} - \text{logit } \pi \right) \\
&= \log \frac{c_I'}{c_I} + \text{logit } \pi + \log \left(\frac{1}{\pi} - \frac{\pi'}{\pi} \right) - \log \frac{\pi'}{\pi}. \quad (4.4)
\end{aligned}$$

The ratios $\frac{c_I'}{c_I}, \left(\frac{c_{II}}{c_{II}'} \right)^{-1}, \frac{\pi'}{\pi}$ can be defined as upper and lower bands on the relative mutability of the initial values. Thereby, one must sustain two properties: $0 < c_I, c_{II}$ and $0 < \pi, \pi' < 1$. Example. The cost ratio appears to be high. Assuming the denominator cost term as $c_{II} = 1$, the other cost term is adjusted to $c_I = 275\,000$. Then, a $\pm 15\%$ mutability band might be presumed, accounting for uncertainty in the definition of the cost ratio. In contrast, the definition of the prior ratio is more certain, however, a $\pm 5\%$ mutability band is set for π (just in case): upper and lower bounds to the threshold result as $5.78 \leq \eta'_{+2} \leq 8.07$ representing $3.08 \times 10^{-3} < \tilde{\pi} < 3.13 \times 10^{-4}$. In other words, a scenario of 1% attacks became an effective scenario of 99.69% to 99.97% attacks—the value of $\tilde{\pi}$ is a mapping of prior and cost beliefs and for revealing the effective prior in the outlined scenario, the cost ratio is simply set to unity $\frac{c_I}{c_{II}} = 1$.

Thus, one may want to start from an operating point representing a verbal band and proceed with a fine-tuning of the decision policy parameters (c_I, c_{II}, π) . In this context of employing mutability ranges,

⁹² The prior is the biggest difficulty in the forensic evaluative scenario.

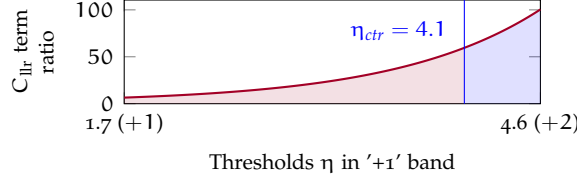


Figure 4.7: Contribution: representative operating points per verbal band, exemplary on the scale of conclusion's +1 band with equal areas left and right to the center of gravity threshold η_{ctr} .

initial thresholds can be denoted, such that calibration performance to either end of a verbal scale band is sustained symmetrically. Therefore, seeking the center of gravity (of costs) is proposed depending on a verbal band. Since DCFs are dependent on $\tilde{\pi}$ and different DCF setups are compared, an application-independent cost measure is necessary. Thus, C_{IIr} is used, see section 2.4 and Eq. (2.54). C_{IIr} terms are examined depending on the ratio of the class \mathcal{A} and class \mathcal{B} score sets $S_{\mathcal{A}}, S_{\mathcal{B}}$. Emphasis on security and convenience scenarios, therein, is symmetrically from an operating point central within a verbal band, i.e., the integral of C_{IIr}^{ratio} :

$$C_{IIr}^{ratio}(\eta) = \frac{\log(1 + e^{a\eta})}{\log(1 + e^{-a\eta})} \quad \text{with} \quad a = \text{sign}(\eta). \quad (4.5)$$

Fig. 4.7 illustrates the determination of such a representative point for the scale of conclusion's +1 band.

Thereby, the EER line is resembled for $\eta = 0$ on well-calibrated systems, cf. Fig. 4.5. At $\eta = 0$, prior and cost beliefs are in equal ratio ($c_I = c_{II} = 1, \pi = 0.5$) but can also be under full uncertainty—the first notion is a throughout parameterization of BDF parameters; the second notion is derived from information theory (the BDF accommodates for both perspectives). By reaching towards higher levels of security or convenience, the centers collapse towards the outer limits of the depending verbal band due to the increasing exponential cost penalty, cf. Tab. 4.2, exemplarily depicting centers of gravity for the scale of conclusion.

To derive operating points verbally, [biometric system vendors](#), [operators](#), [providers](#), and [owners](#) can first discuss the application type, e.g., in terms of $-3, \dots, +3$ verbal bands. Second, they can agree on a range of considerable priors⁹³, then derive dependent costs, cf. Eq. (2.37). Third, they adjust the threshold depending on the representative operating point by utilizing Eq. (4.4), e.g., the cost ratio proposed by the representative operating point of a scale may vary in a $(-10\%, +15\%)$ band or else needs to be downscaled to a distinct cost ratio. By adjusting thresholds, other verbal bands can be effectively reached, e.g., a threshold of scale +2 might increase to scale +3 (which is just fine as the beliefs in a decision policy are updated).

⁹³ One may interpret $1 - \pi$ as the prior 'attack probability' to a system.

4.3 CONTRIBUTION: BINARY-DECISION ERROR TRADE-OFF (BET) PLOTS

In this section⁹⁴, an error rate trade-off is proposed *exactly* accounting for the BDF paramount to decision making. Common ways of measuring the trade-off between error types are ROC and DET [88] plots, where systems are tested for any possible threshold (application). The Type I and Type II errors of a system are plotted as linked pairs, the so-called *operating points*. DET plots are a variant of ROC plots, which plot Gaussian-warped axes.⁹⁵ This variant has two advantages: (i) Gaussian-distributed scores generate straight lines in DET plots; (ii) the low-error region of the graph is magnified, helping to visualize differences in low-error systems. Although theoretically appealing, the Gaussian warping of the axes is mainly motivated because many score normalization techniques tend to Gaussianize the scores [89]. However, this does not take score calibration into account, and therefore systems presenting good DET plots might lead to bad decisions under BDF.

Here, the **binary decision error trade-off (BET)** plot is proposed, an alternative warping of the ROC axes to support the visualization of distances between pairs of Type I and Type II errors, cf. Fig. 4.8, involving three main processes:

1. The first stage involves warping the axes of the ROC plot according to the log-odds transformation, i.e., the *logit* function, instead of the *probit* function of the DET. Thus, an equivalent increase or decrease of the application Bayes threshold towards any of the extreme application scenarios will be visualized as an equivalent increase or decrease in the axes of a BET curve.
2. The second stage is a necessary optimal score calibration. The previous property is manifested only if the scores are optimally-calibrated, which is achieved by using the **pool adjacent violators algorithm (PAV)** [78, 90]. Therefore, the previous visualization property in BET plots is possible.
3. The third stage encodes system curves with scales of decision policies. This is a consequence of the second stage: the **ROCCH** corresponds to the PAV output [78, 90]. LLR thresholds are encoded on the ROCCH in scales of *least-favorable decision policies*.⁹⁶ This way, a more pessimistic evaluation of performance in the empirical set is sought, which is extrapolated as a prediction of future operational performance.

⁹⁴ Parts of this section are based on a collaborative work with Daniel Ramos and Didier Meuwly [65].

⁹⁵ Axes are warped by the *probit* function, whose inverse is the cumulative distribution function (cdf) of the standard normal distribution.

⁹⁶ When \mathcal{A} is chosen over \mathcal{B} as an LLR equals the threshold of a policy.

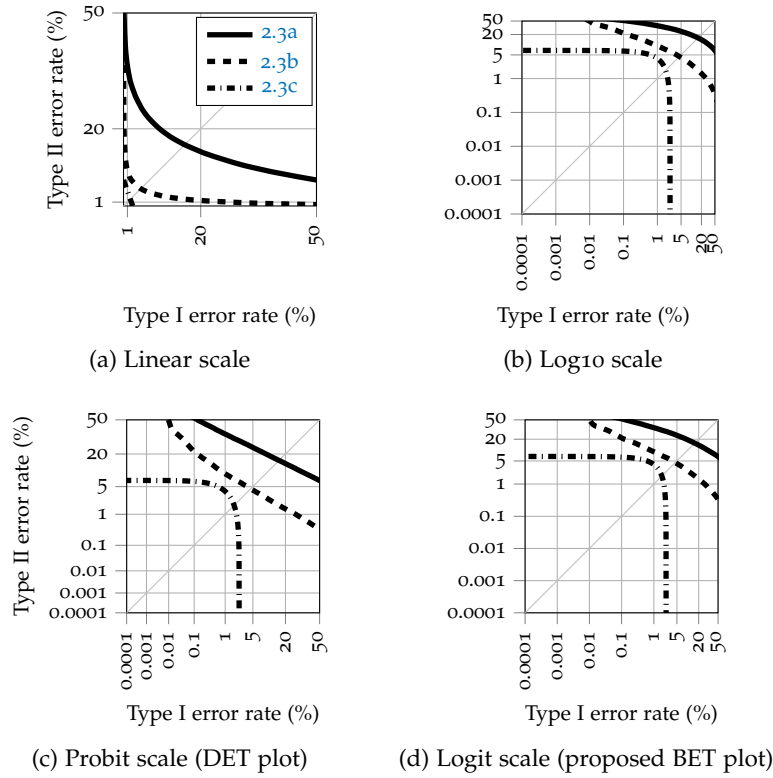


Figure 4.8: Performance visualizations on synthetic score sets (red versus green) sampled from 2.3a $\mathcal{N}(0, 1)$ versus $\mathcal{N}(3, 2)$ (solid), 2.3b $\chi^2(3)$ versus $25 \text{ Beta}(2, \frac{1}{2})$ (dashed), 2.3c $\mathcal{U}(-4, \frac{1}{10})$ versus $\text{Beta}(\frac{9}{10}, \frac{1}{2})$ (dash-dotted), and ROCs with different scales in (a) to (d).

In contrast to other members of the ROC plot family, which solely aim to visualize score distributions in one or another way, the BET plot builds upon an implication of the BDF that is well-known among forensic statisticians: *the LLR of the LLR is the LLR* [39, p. 79]. In other words, applying the LLR principle on scores as well as on features leads to the same value. LLRs encode the class proportion within both spaces in their values. Thus, the meaning of LLRs is invariant to whether the score space or the feature space is discussed: LLRs are ideal for formal assessment of decision performance with binary outcome which is informed by a machine learning system, such as biometric verification. In the logarithmic domain, decision trade-offs resemble a *tug of war* between two opposing propositions: an equal change in favoring either proposition means the same; LLRs are the natural choice for binary decision scores. Even if a recognition system is not capable of yielding LLRs as scores, the second stage of the BET plot will empirically exploit depending ideal (oracle) LLR values by carrying out an oracle *score calibration*. Thus, the latent decision subspace is revealed, the only domain where formal decision models can be applied in a meaningful manner by means of the BDF. Fig. 4.9 shows the BET stages at a glance.

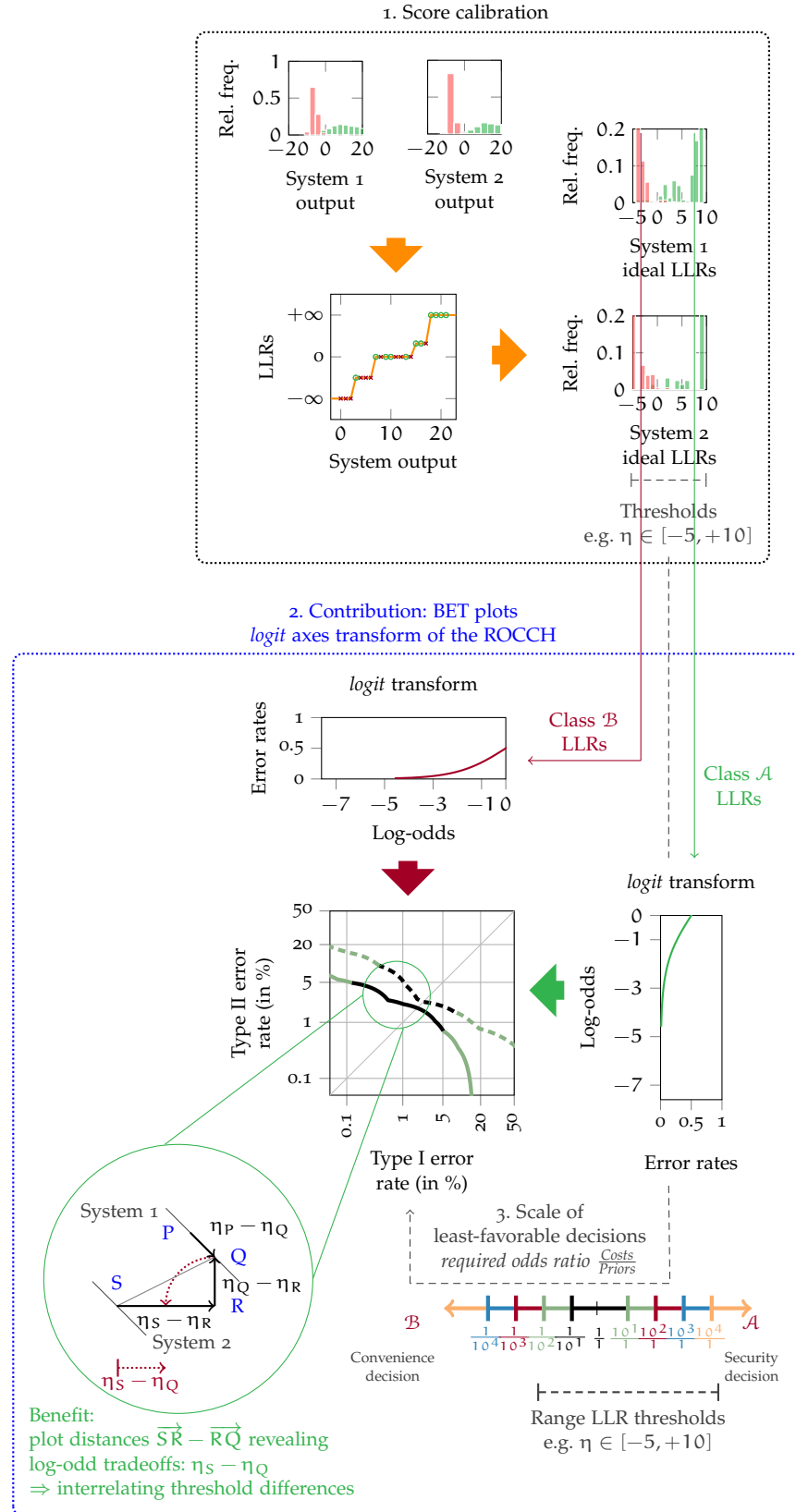


Figure 4.9: Big picture: the BET plot, interrelating Bayes threshold distances, and the empirical support of decision scales.

4.3.1 Score Calibration: Properties and Implications

Score calibration is a necessity for decision making which is informed by machine learning, when class discrimination and the preservation of an ideal decision boundary are intended. In *decision theory*, the accuracy of (probabilistic) predictions is addressed by *proper scoring rules* [78, 96–98], with Brier [96] contributing the most prominent paper. For binary classifiers, however, C_{llr} [26] is the *de facto* standard, a proper scoring rule for LLRs, which emerged in the speaker recognition community. Its basis is rooted in decision theory for pattern recognition [56]: optimal decisions minimize the expected posterior risk (cost) $R(\mathcal{P} | E)$ of favoring proposition \mathcal{P} in the light of some evidence E and a true proposition \mathcal{T}_i with i as proposition index for $N = 2$ propositions:

$$R(\mathcal{P} | E) = \sum_i C_{\mathcal{P}, \mathcal{T}_i} \Pr(\mathcal{T}_i | E) \quad (4.6)$$

with the non-negative costs $C_{\mathcal{P}, \mathcal{T}_i}$ of deciding for proposition \mathcal{P} in the light of the true proposition \mathcal{T}_i . For binary decisions, only the propositions \mathcal{A}, \mathcal{B} are possible for \mathcal{P} and \mathcal{T}_i . Here, decisions are made based on the extent to which proposition \mathcal{A} is favored over proposition \mathcal{B} . The tie-breaking rule to minimize $R(\mathcal{A} | E)$ [56] is (here written down as a least-favorable decision; \leq instead of $<$ as in [56]):

$$\begin{aligned} R(\mathcal{A} | E) \leq R(\mathcal{B} | E), \quad \frac{C_{\mathcal{A}, \mathcal{B}} - C_{\mathcal{B}, \mathcal{B}}}{C_{\mathcal{B}, \mathcal{A}} - C_{\mathcal{A}, \mathcal{A}}} &\leq \frac{\Pr(\mathcal{A} | E)}{\Pr(\mathcal{B} | E)}, \\ \log \frac{C_{\mathcal{A}, \mathcal{B}} - C_{\mathcal{B}, \mathcal{B}}}{C_{\mathcal{B}, \mathcal{A}} - C_{\mathcal{A}, \mathcal{A}}} \frac{\Pr(\mathcal{B})}{\Pr(\mathcal{A})} &\leq \log \frac{\Pr(E | \mathcal{A})}{\Pr(E | \mathcal{B})}. \end{aligned} \quad (4.7)$$

The last line makes use of *Bayes' theorem*, putting priors and costs on the left-hand side, applying the logarithm to both sides (prior and cost terms must be positive, otherwise a recognition task would be superfluous), leaving the LLR to remain on the right-hand side. This very step foreshadows the main difference to the perception of performance visualization established in machine learning, e.g. [203, 204], which operates on posteriors (comparing them to costs): only when a classifier informs on the strength-of-evidence by means of LLRs, evidence is reported comparable to the required trade-off formalized by quantified prior and cost beliefs; the *subjective* decision layer is formally (and entirely) decoupled from *objective* evidence reporting layer, as required in forensic evidence assessment [23, 24]. For the sake of easier tractability, zero costs in making correct decisions $C_{\mathcal{A}, \mathcal{A}} = 0 = C_{\mathcal{B}, \mathcal{B}}$ are assumed. Furthermore, the notation of costs on making an erroneous decision is simplified as: $C_{\mathcal{A}} = C_{\mathcal{B}, \mathcal{A}}$ if \mathcal{A} is true but the decision is \mathcal{B} ; and $C_{\mathcal{B}} = C_{\mathcal{A}, \mathcal{B}}$ if \mathcal{B} is true but the decision is \mathcal{A} . For well-calibrated scores, the Bayes threshold η is, cf. Eq. (2.33):

$$\eta = \log \frac{C_{\mathcal{B}} \Pr(\mathcal{B})}{C_{\mathcal{A}} \Pr(\mathcal{A})} = \text{logit}(1 - c) + \text{logit}(1 - \pi). \quad (4.8)$$

Decision policies are formally denoted by quantifying the belief in priors and costs; *score calibration* means to meet these trade-off requirements with the output of the machine learning systems (not only to improve on *discrimination* but also to sustain *interpretation* in terms of the left-hand side), which informs the decision layer, e.g. a biometric verification system (regardless of the biometric modality being voice, iris, handwriting, fingerprint, face, or another; and regardless of the shape of any empirical score distribution).

4.3.2 Modeling Bayesian Decision Policies (not Scores)

The proposed BET plot is motivated by the visual suggestion of ROC and DET plots: trade-offs in decision policy parameterizations (c, π) are visualized in terms of their error rates, namely Type I error rate in the x-axis and Type II error rate in the y-axis for a given value of threshold η . In BET plots, the axis warping is defined by a quantile function intrinsically given by η . Thus, same distances in the axes reveal equivalent variations in the application policies. Moreover, each point of a BET plot will reveal a pair of Type I and Type II errors that are optimal for the given Bayes threshold. Thereby, the threshold is meaningfully moved across the BET characteristic, considering that a move that supposes an equal distance interrelates to an equivalent increase or decrease in the parameterized decision policy (c, π) .

The threshold η and decision policy parameters (c, π) relate with the *logit* function, which is here proposed as the ROC axes warping. Adopting the logistic distribution⁹⁷, whose cdf is the *logistic* function, cf. Eq. (2.34), and whose quantile function is the *generalized logit* function, the integrals of Type I and Type II error rates p_1, p_2 are sampled by its cdf, and corresponding thresholds η_1, η_2 are modeled by its quantile function:

$$\eta_1 \sim \mu_1 + s_1 \logit p_1, \quad \eta_2 \sim \mu_2 + s_2 \logit p_2 \quad (4.9)$$

with location and scale parameters for the Type I error rate μ_1, s_1 and the Type II error rate μ_2, s_2 .

Theorem. *By exploiting the latent decision subspace, LLRs (having an angular interpretation in the ROC canvases) are co-assigned with the vertical/horizontal linear interpretation (which the human eye is used to judge on in error rate trade-off plots).*

Due to its underlying decision model, the BET plot is the first error rate trade-off plot to fully interrelate error rate and Bayesian performance assessment paradigms *precisely*. In the following, proofs are outlined, implications are stated, and examples are provided.

⁹⁷ As Gaussian-distributed scores resemble straight lines in the DET space, logistically distributed scores will resemble straight lines in the BET space, but first and foremost, the BDF is targeted, i.e., the decision formalism.

4.3.3 Tractable Decision Model Parameters

By definition of calibration, calibrated probabilities are equal to error rates (on average). The logistic quantile distribution of the threshold η thus naturally appears as the logistic function. The relationship between LLRs and posterior probabilities may be viewed as a generalized sigmoid. Thereby, coordinates after PAV (p_1, p_2) are examined, i.e., points of the ROCCH that are interpreted as perfectly-calibrated posteriors. Employing s_1, s_2 scalings, competing log-odds are weighted. For the sake of easier tractability, let $s_1 = s_2 = 1$, such that solely μ_1, μ_2 terms are discussed in the following.

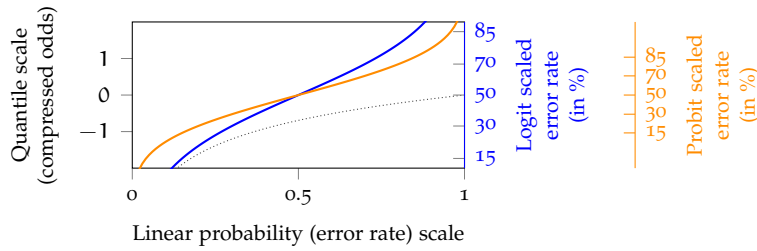


Figure 4.10: Comparison of *logit* (blue) and *probit* (orange) scales, reflecting probabilities as compressed odds. The *log* transform (dotted) is solely capable of compression: log-scaled error rates cannot warp to positive values, whereas compressed odds are symmetric in $(-\infty, +\infty)$ towards 50%.

Fig. 4.10 compares the *log*, *probit* (DET), and *logit* (proposed BET) axes transform functions. By employing the *logit* axes scaling, distances in the BET space account for distances between LLR thresholds. In other words, the BET plot provides an *exact* visualization of trade-offs in decision making, as competing (betting) odds are reflected in a symmetric fashion (in terms of log-odds). In contrast, DET plots reflect probabilistic odds, i.e., putting decision making in context of hypotheses testing, whereas the BET plots put decision making in context of favoring propositions based on the strength of evidence. Also, *log*-scaled axes are undesirable: by solely reflecting the odds of one proposition, the odds of the competing proposition are missed out on. In other words, the approximation of the *logit* with the *log* axes scaling introduces the non-linear approximation error: $\text{logit}(x) - \log(x) = -\log(1-x)$, such that visual distances between error trade-offs cannot be *exactly* interrelated with LLR threshold distances. When consequently following the BDF, the *logit* scale is preferable over the *probit* scale, i.e., the BET plot over the DET plot, as log-odds are visually interrelated for decision making.

Fig. 4.11a depicts an exemplary BET plot, cf. Fig. 4.8d, with a colored verbal scale (e.g., of the ENFSI scale or the scale of *least-favorable decisions*), cf. Fig. 4.4a. Evaluation transparency is enhanced as threshold values and distances are interrelated.

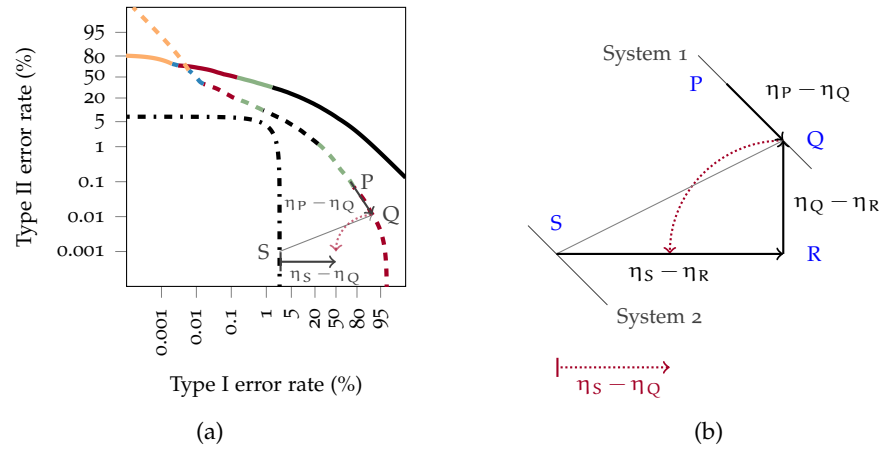


Figure 4.11: (a) BET plot with exemplary systems of Fig. 2.3a (solid), 2.3b (dashed), 2.3c (dash-dotted) and verbal scales of *least-favorable decisions*, (b) interrelated distances between *operating points* S, P, Q, R with the related thresholds $\eta_S, \eta_P, \eta_Q, \eta_R$.

4.3.4 Plot Distances Revealing Log-Odd Trade-Offs

The BET plot interrelates log-odd (c, π) trade-offs among points of one system's ROCCH but also among different systems. Thereby, priors and costs are related in the same way: a variation in priors will cause the same effect as an alike variation of costs. For this reason and for the sake of easier tractability in the following discussion, let $\pi = 0.5$ (logit $\pi = 0$), e.g., one might assume *maximum prior entropy* (no information).⁹⁸ As the scores to be evaluated by the BET plot are PAV-transformed, the resulting plot represents the ROCCH of the scores evaluated [90], i.e., the performance of *least-favorable decisions*. This convexity assessment of *least-favorable decisions* leads to thresholds parameterizations by ratios (that are up to scaling factors), i.e., pessimistic thresholds on the ROCCH (continuous thresholds) instead of on the ROC (discrete thresholds), as follows.

The quantile p_2 represents the thresholds $\eta_2 = \eta_A$, i.e., computed over class \mathcal{A} similarity. Thus, the quantile function reveals a *Type II cost* $c_{II} \models c$ associated to the depending η_A with the competing *Type I cost* $c_I \models (1 - c)$. Inherently, these $\frac{c_I}{c_{II}}$ ratios are up to scaling factors. The costs terms c_I, c_{II} range from 0 to 1, covering all possible cost-ratios and are therefore all possible thresholds for the maximum entropy prior.⁹⁹ By targeting the prediction of decision policy trade-offs,

⁹⁸ The prior $\pi = 0.5$ is the *implicit prior* in all forensic evaluative benchmark tests. Although the value 0.5 is almost never met in forensic science evaluation, the value is motivated by *maximum uncertainty* from information theory. In contrast to the EER (Frequentist perspective on equidistant operating points), however, prior and cost beliefs of 0.5 resemble at the value of maximum *Shannon entropy* (Bayesian perspective on information theory).

⁹⁹ Following Eq. (2.33), when c moves from 0 to 1, η moves from $-\infty$ to $+\infty$ regardless of the (arbitrary but fixed) value of π in $(0, 1)$.

all notations are intended in the latent decision subspace, sparing the assignment of distinct policy parameterization values. When solely interrelating Type II parameters with p_2 , defining μ_2 as (*corollary*):

$$\begin{aligned}\mu_2 &= \log c_I + \text{logit}(1 - \pi), \quad \text{s.t.:} \quad \eta_A \sim \mu_2 + \text{logit } p_2, \\ \Rightarrow \quad \log \frac{c_I}{c_{II}} + \text{logit}(1 - \pi) &\sim \log c_I + \text{logit}(1 - \pi) + \text{logit } p_2, \\ \Rightarrow \quad \text{logit } p_2 &\sim \log \frac{1}{c_{II}}, \quad \text{s.t.:} \quad c_{II} \sim \frac{1 - p_2}{p_2}.\end{aligned}\quad (4.10)$$

Proof: The Type II error cost c_{II} is modelled only by the odds representation of Type II error proportions; any *vertical* change in the BET canvas corresponds to a change in c_{II} costs, which in turn reflects an LLR threshold deviation from its above expectation η_A .

By contrast, the quantile p_1 represents the threshold $\eta_2 = \eta_B$, i.e., the quantile of a right-to-left integral computed over class \mathcal{B} similarity, inverting LLR thresholds η_A : $\eta_B = -\eta_A$. When solely interrelating Type I costs with p_1 , let μ_1 (*corollary*):

$$\begin{aligned}\mu_1 &= \log c_{II} + \text{logit } \pi, \quad \text{s.t.:} \quad \eta_B = -\eta_A \sim \mu_1 + \text{logit } p_1, \\ \Rightarrow \quad -\log \frac{c_I}{c_{II}} + \text{logit } \pi &\sim \log c_{II} + \text{logit } \pi + \text{logit } p_1, \\ \Rightarrow \quad \text{logit } p_1 &\sim \log \frac{1}{c_I}, \quad \text{s.t.:} \quad c_I \sim \frac{1 - p_1}{p_1}.\end{aligned}\quad (4.11)$$

Proof: The Type I error cost c_I is modelled only by the odds representation of Type I error proportions; any *horizontal* change in the BET canvas corresponds to a change in c_I costs, which in turn reflects a LLR threshold deviation from its above expectation η_B .

Similarly to c_{II} , c_I increases by moving towards the origin of the BET plot and decreases when the depending error rate approaches 100%. Thereby, thresholds η ($\eta_1 = -\eta_2$) increase either when c_I increases or c_{II} decreases. Solely traversing on the x - or y -axis from a threshold η' to a threshold more close to the BET origin η'' —for $p_1'' < p_1' : \eta'' > \eta'$, whereas for $p_2'' < p_2' : \eta'' < \eta'$ —distances resemble (*corollary*):

$$\begin{aligned}\text{x-axis: } \text{logit } p_1'' - \text{logit } p_1' &\sim \log c_I' - \log c_I'' = \eta'' - \eta', \\ \text{y-axis: } \text{logit } p_2'' - \text{logit } p_2' &\sim \log c_{II}' - \log c_{II}'' = \eta' - \eta''.\end{aligned}\quad (4.12)$$

Proof: Changing horizontally from lower to higher Type I error rates corresponds to (the distance when) changing from more to less secure thresholds; changing vertically from higher to lower Type II error rates corresponds to (the distance when) changing from more to less convenient thresholds (from less to more secure thresholds).

Fig. 4.11b provides a visual intuition on directed distances in the BET canvas, where points P, Q are on the same ROCCH, S is on a different, and R is an auxiliary point, necessary to interrelate the threshold difference from Q to S.

4.3.5 Generalization to Priors as well as Priors and Costs

Solely examining priors ($\text{logit } c = 0$), the parameterization of μ_1, μ_2 reveals *latent priors* $\pi_{II} \models \pi$ and $(1 - \pi_I) \models (1 - \pi)$ (up to scaling factors), where expected thresholds for one decision model the expected parameters of the competing decision (similarly to the above):

$$\begin{aligned} \mu_1 &= \log \pi_{II} + \text{logit}(c), \quad \text{s.t.:} \quad \text{logit } p_1 \sim \log \frac{1}{\pi_I}, \\ \mu_2 &= \log \pi_I + \text{logit}(1 - c), \quad \text{s.t.:} \quad \text{logit } p_2 \sim \log \frac{1}{\pi_{II}}. \end{aligned} \quad (4.13)$$

Corollaries and *proofs* correspond to the above.

Due to the threshold symmetry regarding cost and prior log-odds, cf. Eq. (2.33), the above argumentation holds for solely discussing different prior parameterizations as well as for (c, π) as (*corollary*):

$$\begin{aligned} \text{logit } p_1 &\sim -\log c_I - \log \pi_I + \log c_{II} + \log \pi_{II} - \mu_1, \\ \text{logit } p_2 &\sim -\log c_{II} - \log \pi_{II} + \log c_I + \log \pi_I - \mu_2, \quad \text{where if} \\ \mu_1 &= \log c_{II} + \log \pi_{II} \quad \Rightarrow \quad \text{logit } p_1 \sim -\log(c_I \pi_I), \\ \mu_2 &= \log c_I + \log \pi_I \quad \Rightarrow \quad \text{logit } p_2 \sim -\log(c_{II} \pi_{II}). \end{aligned} \quad (4.14)$$

Proof: Type I error rates interrelate with the convolution of depending latent priors and costs π_I, c_I ; Type II error rates interrelate with the convolution of depending latent priors and costs π_{II}, c_{II} . Analogous to what is stated above, vertical and horizontal LLR threshold changes are impacted accordingly.

The illustrated proofs work out well for the proposed formalized decision modeling on the BET canvas because of the fact that LLRs are not only scores of some form to report on the class distribution within the feature space but also encode in their value the class proportion within the latent decision subspace: *the LLR of the LLR is the LLR*.

4.3.6 Implications for Interrelations

The BET model allows to interrelate a formal interpretation of fix error rate constraints, exemplary the constraints by FRONTEx for face biometrics in automated border control [77]: if p_2 performances are compared at a fix $p_1 = 0.1\%$ error rate, the latent cost $c_I = \frac{1-10^{-3}}{10^{-3}} = 999$ is induced. Then, by requiring $p_2 \leq 5\%$, $c_{II} \geq \frac{1-0.05}{0.05} = 19$ and $\eta \leq \log \frac{999}{19} \approx 3.96$ follow.¹⁰⁰ In other words, convenience is increased at a fixed security level, accommodating for lower Bayes thresholds as well. The BET model allows for shifts between beliefs in the decision policy while sustaining error rate constraints. For example, the above assumed maximum prior uncertainty ($\pi = 0.5$) for a cost belief of $c \geq \frac{19}{19+999} \approx 0.019$. The application environment is shifted to

¹⁰⁰ Comparatively, the 2016 NIST speaker recognition evaluation [22] sets its lowest operating point at $\eta_{\text{SRE}_{16a}} = \text{logit}(1 - 0.01) \approx 4.60$.

be more friendly with $\pi = 95\% = \frac{19}{20}$, where $c_{II} \geq 19$ is kept alongside the error rate constraints $p_1 = 0.1\%$ ($\text{logit } p_1 = -\log(999)$) and $p_2 \leq 5\%$ ($\text{logit } p_2 \leq -\log(19)$):

$$\begin{aligned} \text{logit } p_2 &\sim -\log(c_{II} \pi_{II}) \Rightarrow -\log(19) \sim -\log(19 \pi_{II}) \Rightarrow \pi_{II} \sim 1, \\ \frac{\pi_{II}}{\pi_{II} + \pi_I} &\models \frac{\pi}{\pi + (1 - \pi)} = \pi \Rightarrow \pi_I = \frac{1 - \pi}{\pi} = \frac{1}{19}, \\ \text{logit } p_1 &\sim -\log(c_I \pi_I) \Rightarrow -\log(999) \sim -\log\left(c_I \frac{1}{19}\right) \Rightarrow c_I \sim 999 \times 19, \\ \frac{c_{II}}{c_{II} + c_I} &\models \frac{c}{c + (1 - c)} = c \Rightarrow c \geq \frac{19}{19 + 999 \times 19} = \frac{1}{1000}. \quad (4.15) \end{aligned}$$

The mutual prior and cost beliefs of a decision policy (c, π) can be interrelated with fixed to error rate constraints. By changing one of these beliefs (e.g. π), however, the other belief (e.g. c) is affected. Thus, in the example, the cost term of an application's decision policy changes from a rather convenience centered yet secure ($c \rightarrow 1$, here: $\frac{19}{1018} \approx \frac{1}{54}$) to a much more security centered one ($c \rightarrow 0$, here: $\frac{1}{1000}$).

4.3.7 Examples: Experimental and Use-Case

The proposed BET plots can be applied to any binary decision problem, as long as the decision is made by the use of a score as compared to a threshold. In this section, two examples are discussed: i) advantages of the BET plot on scores stemming from an automated fingerprint identification system (AFIS)¹⁰¹ in forensic applications and ii) utility of the BET plot to smart home applications.

Example: AFIS System (Real-World) I/II

AFIS systems use fingerprint papillary line features (e.g., minutiae) as biometric characteristics for the purpose of biometric recognition, e.g., verification. Thereby, the number of minutiae examined by an AFIS system correlates with the discrimination and calibration performance [76, 205]: performance metrics improve with the increasing number of minutiae. Fig. 4.12 compares AFIS scores^a, cf. [76], to their ideal LLR values: as the score distributions, cf. Figs. 4.12a–4.12d, solely provide insights into associated error rates, when compared to a threshold, their ideal LLR representation, cf. Fig. 4.12e–4.12h, provides insights into decision making because it represents the ratio of probabilities of observing scores given the two competing propositions \mathcal{A}, \mathcal{B} . In [76], results are visualized in terms of DET [88] and ECE [115] plots because both

¹⁰¹ Such systems are primarily used for investigative purposes with the aim of minimizing the false reject rate (FRR), not the false accept rate (FAR). The investigative approach is a selection from larger sets to smaller subsets in which the FRR is minimized. At the evaluation level, however, it is the contrary. Here, the aim is to get the highest LLR (at the right level of inference) so as to minimize the FAR (even if this information remains prior dependent).

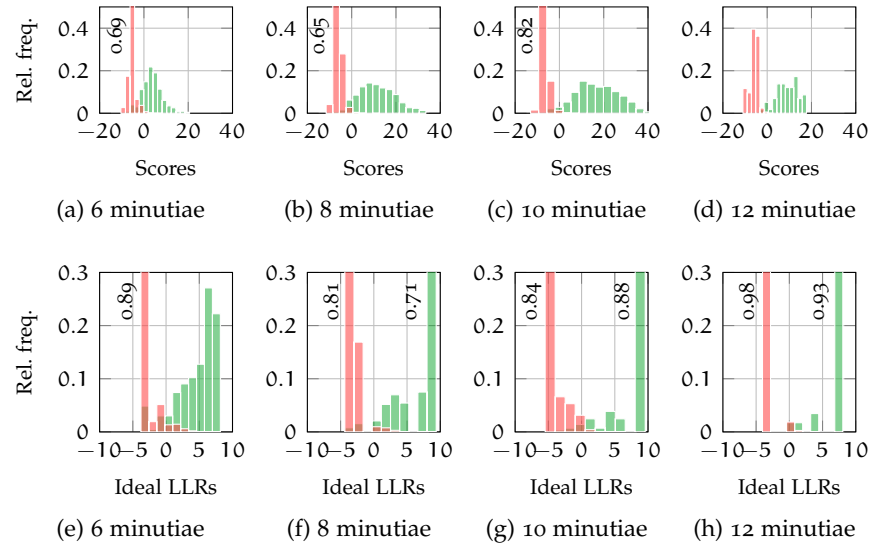


Figure 4.12: BET plot example: AFIS performance by number of examined minutiae with (a)–(d) AFIS scores, (e)–(h) related LLRs.

discrimination and calibration are considered for LLRs. However, when solely concerning discrimination performance (as for the BET plot), the examination of calibration performance (by ECE plots) is beyond the scope of the BET plot, such that calibration is assumed to be ideal.

^a AFIS scores have been used secondarily to compute the strength of evidence in this forensic evaluation scenario.

Therefore, ROC plots can solely measure discrimination, and the PAV transformation does not change the ROC. Thus, empirical (AFIS) scores are transformed into their depending LLR values (after conducting PAV), where (LLR) decision trade-offs resemble the ROCCH, cf. Fig. 4.13, when continuously increasing decision thresholds.

Example: AFIS System (Real-World) II/II

As the discrimination performance increases, fewer LLR values resemble after score calibration, such that the ROCCH is defined by fewer points. Consequently, steppey ROC plots, solely discretely sampling thresholds, are not relevant to the BET space. Thresholds are examined continuously and in the LLR domain (on well-calibrated scores) by the use of the ROCCH. Depending on the underlying LLR values, different decision scales are supported. When employing verbal scales of least-favorable decisions, cf. Fig. 4.4a, depending LLR threshold values are also interrelated. When examining 10 minutiae, the threshold scale of *moderate security* decisions (comparing the green BET segments) is rather narrow compared to other numbers of minutiae. By comparing the green BET segments (resembling $\text{LLRs} \in (\log(10), \log(100)] \approx (2.3, 4.6]$), a par-

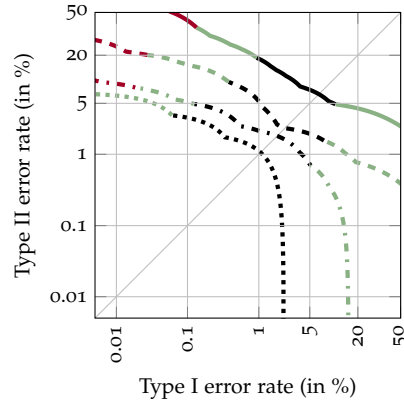


Figure 4.13: BET plot example: AFIS BET plots for 6 minutiae (dashed), 8 minutiae (dotted), 10 minutiae (dash-dotted), and 12 minutiae (solid); colors depict the verbal scale (ENFSI based of *strength of evidence* and of *least-favorable decisions*).

ticular consequence of the transform from scores as system outputs to their corresponding ideal LLR values is revealed: optimal score calibration maps uncalibrated scores to corresponding LLR values—these effective LLR groups are not necessarily continuous, but they are monotonically increasing. Depending on the empirical set of uncalibrated scores, different gaps between calibrated system outputs (as LLRs) can resemble. In PAV calibration, these gaps are linearly interpolated. On the ROCCH, these gaps are angularly interpolated, i.e., at the angles between two neighboring ROCCH segments. Comparing 10 and 12 minutiae LLR score histograms, cf. Figs. 4.12g and 4.12h, the error rates associated to LLRs span over a wider range of *least-favorable decisions* for the 10 minutiae AFIS. Therefore, the 12 minutiae AFIS yields better error rate trade-off performance. After optimal calibration, LLR values group by the empirical ratios of class A, B scores, cf. Fig. 2.12. Depending on the LLR value of and gaps between these groups, narrower and wider ranges resemble. BET plots allow to interrelate (LLR) score histograms of the decision subspace with error rate trade-offs. Furthermore, the changeover between verbal scales is not occurring at fixed error rates: good decision making needs to discuss the formalized operating point (c, π) in order to avoid misleading decisions (on average).

The utility of BET plots is not restricted towards forensic scenarios as the benefit of interrelating betting odds is applicable to many binary decision problems, such as smart home applications.

Example: Smart Home Applications (Use Case)

Speaker verification could be employed for user authentication, accessing lighting control or home security systems. In such environments, one might assume high prior odds $10 : 1$ with $\pi = \frac{10}{11}$ (true

identity claims are ten times more likely than false claims). For non-critical applications, such as lightning control, one might assume high convenience cost odds with $c = \frac{2}{3}$ (e.g., inconvenience versus power consumption costs $C_A = 20$ cents, $C_B = 40$ cents), whereas for critical applications such as home security systems, one might assume high security cost odds with $c = \frac{1}{991}$ (e.g., inconvenience versus liability costs $C_A = 100$ €, $C_B = 99\,000$ €). The first application implies the LLR threshold $\eta = \text{logit}(1 - \frac{20}{21}) = -\log(20) \approx -3.0$ in order to favor proposition A —making at least *moderate convenience* decisions, cf. Fig. 4.4a—the latter application requires LLR threshold $\eta = \text{logit}(1 - 0.01) = \log(99) \approx 4.6$ —which provides at least *moderate security*^a decisions, cf. Fig. 4.4a. Therefore, the BET plot allows to choose the most adequate operating point, directly relating LLR thresholds and errors with prior/cost trade-offs.

^a In Fig. 4.4a, an LR of 99 resembles a *moderate security* decision, whereas an LR of 100 resembles a *moderately strong security* decision. It is obviously understood that an LR of 99 is only trivially different in its overall impact from one of 100 [23] (the quote originally refers to LRs of 999 and a 1001 for two verbal bands overlapping at exactly a 1000, but the message remains the same).

Decision policies (c, π) resemble LLR thresholds which correspond to points of the ROCCH. The proposed BET plot visually interrelates ROCCH distances to LLR threshold distances, thus to the effective domain of decision making—the latent decision subspace—rather than to the system output (observation) domain. System outputs might be uncalibrated, and thus would be without any consideration of severity nor hostility aspects (values of c and π) to decision making, especially to erroneously making decisions. Thus, when defining system operating points solely by error rates, systems might be employed causing misleading decisions when either of the cost or prior constraints change. By formalizing application policies (c, π) , the transparency of empirical evaluations is enhanced for a natural interrelation of error rates and the betting log-odds systems are capable of supporting (in application scenarios with multiple changing operating points).

In contrast to empirical ROC plots solely visualizing error rates *as is*, DET and BET performance visualizations provide predictive insights into future decision making. Although motivated differently, DET plots effectively inform on hypotheses testing. By contrast, BET plots inform on trade-offs between (forensic or commercial) decision policies—trade-offs in LR magnitudes. Contrary to literature in machine learning, e.g., [203, 204], BET plots allow to change priors and costs without assuming that the empirical dataset prior equals the operative application prior $\tilde{\pi} = \pi$. In [203], *cost curves* are proposed: the effective prior $\tilde{\pi}$ (there as *probability cost function*) is interrelated with an *average normalised expected cost* ([applied probability of error \(APE\)](#) plots but with predefined cost values and taking the ROC error rates

as prior probabilities). The work of [204] is in opposition to [206, 207]. The two latter argue that the AUC requires *different probability threshold distributions for different classifiers* and assumes *uniform distribution over the proportion of correct classifications*. The former acknowledges a plethora of evaluation metrics (e.g., based on *strictly proper scoring rules*, including C_{llr}) but then explicitly considers the simplest option: *uniform distribution*. In [204], the Brier score (also known as the mean square error) is related with the AUC, and the assessment of LLRs is stated for future work. The calibration of scores and thresholds is elaborated for a fixed database prior $\hat{\pi}$ (used for training systems/describing evaluation database) to equal the prior π (used for defining decision policies). This can be valid when only the cost term of a decision policy changes and the prior term remains fixed and the evaluation database is well-chosen to replicate the application-dependent prior π requirement. This is *not* valid in other scenarios, where systems need to be well-calibrated across different π values to satisfy the (not only cost dependent) requirements of policies ranging between convenience and security decisions, such as in the forensic and banking scenarios.

Contributing BET plots, the choice of the BDF basis to formal decision making is accounted for. System trade-offs are compared on the ROCCH, representing optimal score calibration, i.e., for all possible combinations of prior probabilities and decision costs, optimal Bayes thresholds are found. Finally, the *logit* axes transform interrelates threshold distances, where the visualization of verbal scales further enhances transparency, regardless of the verbal scales being of *strength of evidence* or of *least-favorable decisions*.

Regarding the placement of axes ticks on DET or BET plots, one might favor to have equidistant ticks. In this dissertation, either trade-off employs axes ticks as follows:

- The maximum is either 50% or 99%.
- Error rates below 1% are in magnitudes of base 10 order, e.g., 0.00001%, 0.0001%, 0.001%, 0.01%, 0.1%, 1%.
- Error rates above 1% are limited to five ticks (up to 99%) on approximately equal distances, such that for the *probit* scale¹⁰² (DET plots) and for the *logit* scale¹⁰³. In (BET plots), the ticks are approximated as: 5%, 20%, 50%, 80%, 95%.

¹⁰² See: [https://www.wolframalpha.com/input/?i=0.5*\(1%2Berf\(\(\(probit\(0.99\)++probit\(0.01\)\)+%2F6+***5B1,2,3,4,5%5D+%2B+probit\(0.01\)\)%2Fsqrt\(2\)\)\)](https://www.wolframalpha.com/input/?i=0.5*(1%2Berf(((probit(0.99)++probit(0.01))+%2F6+***5B1,2,3,4,5%5D+%2B+probit(0.01))%2Fsqrt(2))))

¹⁰³ See: [https://www.wolframalpha.com/input/?i=sigmoid\(\(logit\(0.99\)++logit\(0.01\)\)+%5C%2F6+***5C%5B1,2,3,4,5%5C%5D+%5C%2B+logit\(0.01\)\)](https://www.wolframalpha.com/input/?i=sigmoid((logit(0.99)++logit(0.01))+%5C%2F6+***5C%5B1,2,3,4,5%5C%5D+%5C%2B+logit(0.01)))

4.4 CONTRIBUTION: NORMALIZED EMPIRICAL CROSS-ENTROPY

In contrast to error rate performance, this section¹⁰⁴ addresses information based performance assessment within the BDF. The normalization of APE diagrams, cf. [normalized Bayes error rates \(NBERs\)](#) plots of Fig. 2.8 in Fig. 2.9, motivates the following property. The visual comparability of calibration losses from DCF to minDCF values is preserved across different priors, accounting for changing *default* DCF values across operating points—the default performance is the DCF of a coin tossing system which causes different decision costs depending on the decision policy (c, π) parameterization. For ECE plots [115], the default ECE values solely depend on the prior parameterization $\tilde{\pi}$, thus [NECE](#) plots are proposed here as the ratio to the default entropy H_{default} :

$$\text{NECE} = \frac{\text{ECE}}{H_{\text{default}}},$$

$$H_{\text{default}} = \pi_Q \log_2 \frac{\pi_P}{1 + \pi_P} + (1 - \pi_Q) \log_2 (1 + \pi_P), \quad (4.16)$$

where π_Q denotes the reference prior and π_P denotes a system's prior. To simulate information performance over a range of priors, [115] refers to $\pi_Q = \pi_P$. The concept of normalizing cross-entropy is not new, as it is well-known within the [automatic speech recognition \(ASR\)](#) community [208, 209]. Fig. 4.14 depicts the NECEs corresponding to Fig. 2.16. In contrast to ECE plots, visual distances across different prior log-odds (x-axis depicts $\text{logit } \pi$) are comparable, as equal distances reflect the same normalized discrimination and calibration performance.

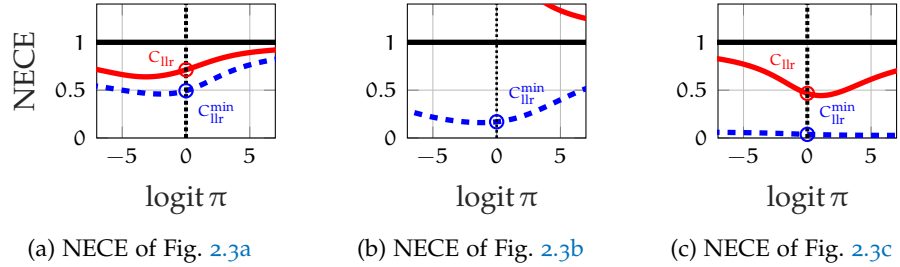


Figure 4.14: Proposed NECE plots for the systems of Figs. 2.3a to 2.3c, with persisted C_{llr} and $C_{\text{llr}}^{\text{min}}$ values at $\text{logit } \pi = 0$ (dashed), default cross-entropy (black), normalized ECE (red), and normalized minECE (dashed, blue).

Notably, the visualization of C_{llr} and $C_{\text{llr}}^{\text{min}}$ values is preserved during ECE normalization. In ECE plots, C_{llr} and $C_{\text{llr}}^{\text{min}}$ values are readable on the ECE characteristics at $\text{logit } \pi = 0$ [115] and the default

¹⁰⁴ Parts of this section are based on a collaborative work with Daniel Ramos and Didier Meuwly [65]. Neither this entire section nor parts of it have been published by the time of the submission of this dissertation.

ECE value.¹⁰⁵ The values of the ECE characteristic at $\text{logit } \pi = 0$ remain invariant to the NECE transform as the default entropy equals 1, i.e., fix-points, such that their values are still readable at the vertical (dotted) line at $\text{logit } \pi = 0$. For the alignment of score histograms and NECE plots, one might intend to flip the x-axis. However, one would thereby visually suggest $\pi = \tilde{\pi}$, which solely holds on $\text{logit } c = 0$ (maximum cost entropy).

4.5 CONTRIBUTION: TAXONOMY TO PERFORMANCE VISUALIZATIONS

This section¹⁰⁶ provides a taxonomy as an overview on the performance visualization approaches depicted in sections 2.2 and 2.4. The plots proposed in sections 4.3 and 4.4 are put into the context of existing work. Performance assessment concepts serve different purposes:

- Solely error rate based analyses are well-established and work for application scenarios with a fixed operating point. When examining different operating points, however, e.g., when researchers or developers interpret published results for their operating scenarios, the impact of employing a system to decision making needs to be considered as well. Two performance reporting types are identified: *discrimination* (class differences) and *discrimination and calibration* (the ability to instrumentalize the knowledge provided by a system to proceed an action).
- Depending on a system's application scenario, different perspectives on performance assessment are applicable, i.e., error rates are at-hand for decision risk assessment. For application-independent reporting, however, a recognition system is thought of as the *weight of evidence* that is contributed to an informed decision making, thus its information gain is assessed. Two performance reporting scopes are identified: *error rates* (whose sets occur as comparison scores) and *information gains* (the knowledge one learns about score sets to provide as beneficial intel to others).
- Overall, the targeted audience needs to be addressed well. APE or ECE plots might report sufficiently enough for analytic purposes in order to develop a system, but NBER and NECE plots ease the visualization for the sake of deciding on which sys-

¹⁰⁵ Regarding the ECE plot, the visualized default value is computed so as to particularly aid the visual assessment. At $\text{logit } \pi = 0$, the default performance evaluates to exactly 1, see: [https://www.wolframalpha.com/input/?i=-\(\text{sigmoid}\(x\) * \log\(\text{sigmoid}\(x\)\) + \(1 - \text{sigmoid}\(x\)\) * \log\(1 - \text{sigmoid}\(x\)\)\) / \log\(2\)](https://www.wolframalpha.com/input/?i=-(\text{sigmoid}(x) * \log(\text{sigmoid}(x)) + (1 - \text{sigmoid}(x)) * \log(1 - \text{sigmoid}(x))) / \log(2)).

¹⁰⁶ Parts of this section are based on a collaborative work with Daniel Ramos and Didier Meuwly [65]. Neither this entire section nor parts of it have been published by the time of the submission of this dissertation.

tem to employ better. As such, steppy ROCs can provide insights into the development of binary decision systems, whereas the contribution to decision making is better reflected by the ROCCH. Two targeted audiences are identified: *analytics* (the examination of performance elements from observations) and *reporting* (to account as a formal statement on observations and investigations).

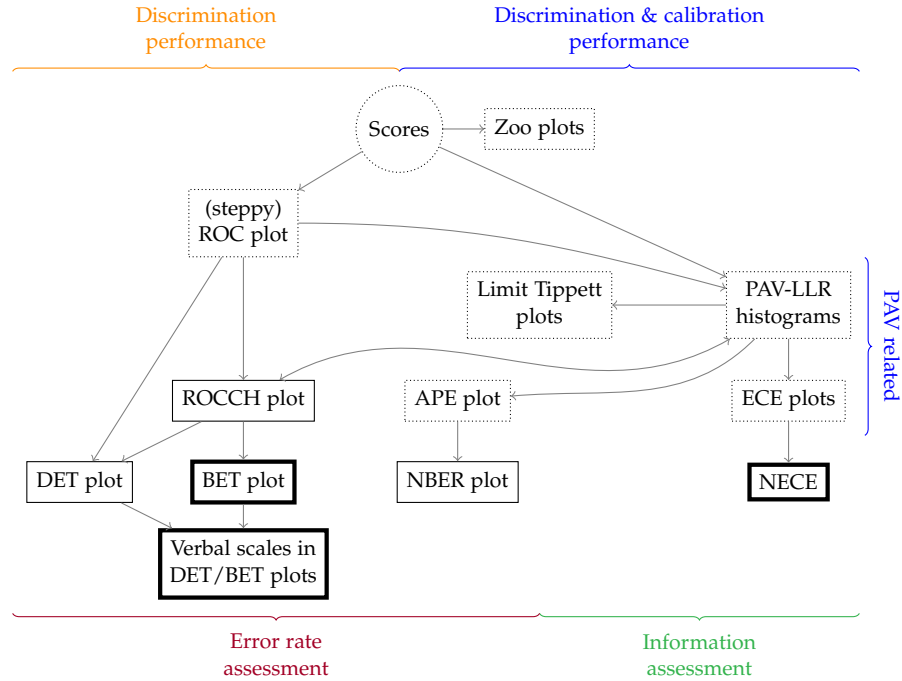


Figure 4.15: Taxonomy on performance visualizations with concepts on: performance reporting types as discrimination (orange) and discrimination and calibration (blue); performance reporting scopes as error rates (red) and information gains (green); and targeted audience as analytics (dotted boxes) and reporting (solid boxes). Plots involving PAV calibration to simulate well-calibrated scores are grouped. Relationships are indicated by arrows (*can be derived from*), but transitive relations of alike reason are omitted. Scores (circle) are the root of this taxonomy. Performance visualizations proposed in this dissertation are highlighted (thick boxes).

Fig. 4.15 gives an overview on well-established performance visualizations of the associated communities, namely speaker recognition, forensic evaluation, and biometrics standardization. The depicted plots relate as follows:

- **Score histograms** visualize discrimination and calibration performance of the system output for the analytic interpretation of score distributions and error rates.

- **PAV-LLR score histograms** visualize inferred information from score histograms on discrimination and calibration performance on the analysis of LLR score distributions and the latent decision subspace.
- **Zoo plots** visualize error rates as well as inferred information from score histograms on the analysis of subject-wise contributions to discrimination and calibration performance. Zoo plots assume scores to be Gaussian distributed for each subject.
- **Limit Tippett plots** visualize error rates as well as inferred information from score histograms on the analysis of the discrimination and calibration performance of system outputs. Thereby, error rate trade-offs are visualized for LLR thresholds (after PAV calibration; thus the transitive dependency to *scores* is omitted). Assumptions on error rate limits are drawn from Gaussian assumptions based on the properties posed by the setup of an empirical evaluation (the numbers of class \mathcal{A} , \mathcal{B} scores). As such, Limit Tippett plots are derivable during the creation of PAV-LLR histograms.
- **(Y-axis inverted, steppy) ROC plots** visualize error rate trade-offs for the analysis of discrimination performance. ROCs are derived from score sets. As ROCs convey error trade-offs, i.e., proportions of score sets, they can serve as input to PAV calibration as well.
- **(Y-axis inverted) ROCCH plots** visualize (a) the convex hull of the ROC and (b) the continuous error rate trade-offs that are sampled after PAV calibration. As such, discrimination performance (representing minDCF values, thus C_{llr}^{\min}) is illustrated for performance reporting.¹⁰⁷
- **APE plots** visualize discrimination and calibration performance (DCF and minDCF values) as well as the default decision risk based on error rates, providing an analysis across all LLR thresholds. The DCF and minDCF integrals equal C_{llr} and C_{llr}^{\min} . To depict calibration performance, APE plots are based on PAV calibration (PAV-LLR histograms).
- **NBER plots** normalize APE plots for the reporting of error rate based discrimination and calibration performance. DCF and minDCF values are normalized by the default decision risk that depends on the parameterization of the *effective prior/cost* $\tilde{\pi}$.
- **ECE plots** visualize information performance inferred from score sets after PAV calibration. By simulating different prior

¹⁰⁷ The ROCCH is the set of the expected ROC segments of *pessimistic* and *optimistic* interpolations [90]. According to [210], *all points under the convex hull are non-optimal*.

π values (performance without the consideration of costs, not to be confused with $\tilde{\pi}$), an analysis of minECE and ECE discrimination and calibration performance is illustrated as well as the default cross-entropy. C_{llr} and C_{llr}^{\min} values are readable from the ECE characteristic at $\logit \pi = 0$. To depict calibration performance, ECE plots are based on PAV calibration (PAV-LLR histograms).

- **NECE plots** normalize ECE plots for the reporting of information based discrimination and calibration performance. ECE and minECE values are normalized by the default cross-entropy that depends on the parameterization of the prior π . C_{llr} and C_{llr}^{\min} values are readable from the NECE characteristic at $\logit \pi = 0$. **(NECE plots are a contribution of this work.)**
- **DET plots** visualize error rate trade-offs for reporting on discrimination performance aiming at hypotheses testing. The axes of the y-axis inverted ROC plot are transformed using the *probit* function. Originally, DET plots were proposed as quantile-quantile (Q-Q) plots since Gaussian distributed scores tend to be visualized as straight lines after these axes transform. The probit function depicts error rates not in the form of probabilities but by their corresponding *probabilistic odds*. In the same way, DET plots visualize ROCCHs.
- **BET plots** visualize error rate trade-offs for reporting on discrimination performance aiming at decision making between competing propositions. The axes of the y-axis inverted ROC plot are transformed using the *logit* function. In this dissertation, BET plots are proposed as Q-Q plots to interrelate trade-offs in prior and cost beliefs with trade-offs in error rates. The logit function depicts error rates not in the form of probabilities but by their corresponding *log-odds*. As the BET plot is motivated by the BDF throughout, ROCs cannot be visualized, whereas the visualization of ROCCHs accommodates to interrelate LLR values (scores after PAV calibration). **(BET plots are a contribution of this work.)**
- **Verbal scales in DET/BET plots** visually summarize LLR score and threshold values (via the ROCCH) for a more transparent reporting on discrimination performance of error rate trade-offs. Depending on the evaluation reporting context, these verbal scales are interpreted and designed towards *strength of evidence* or towards *least-favorable decisions*. **(The visualization of verbal scales on the ROCCH and the proposal of verbal scales of least-favorable decisions are contributions of this work.)**

For the purpose of recommending a few performance visualizations, commercial scenarios might employ *verbal-scaled* BET plots in order

to compare the discrimination performance of multiple systems and NBER plots per system in order to report on calibration performance. By contrast, forensic scenarios might employ *verbal-scaled* BET and NECE plots, where system-depending NECE plots reveal the information gain by employing a system into chains of decision inference.

For scalar metrics, (fast) *analytic* performance estimates might consider the linearly sampled *EER* and C_{llr} , whereas *reporting* performance estimates might consider the ROCCH-EER, the *FNMR at a 1% FMR (FMR₁₀₀)* (or alike requirements), the minDCF (e.g., with $\tilde{\pi} = 0.01$ as of the home security system example), the DCF (e.g., $\tilde{\pi} = 0.01$), C_{llr}^{\min} , and C_{llr} . To limit the metrics to a few, C_{llr} and C_{llr}^{\min} are the application-independent objective (as an application cannot be known when measures are specified covering *all* applications).

4.6 SUMMARY

The performance assessment concepts of conventional error rate trade-off based diagrams are interrelated with the Bayesian paradigm. The formal notation and definition of *angular operating points* is fundamental to the major difference between two perspectives: while Frequentists change thresholds by moving between error rate trade-offs, Bayesians change thresholds by moving between LLRs that, in terms of error rate tradoffs, correspond by changing between segments of the ROCCH, i.e., between the angles of neighboring segments. Based on the ROCCH, bands of LLR scores and thresholds are summarizable by *verbal scales*—verbal scales for scores are referred to as *strength of evidence* and verbal scales for thresholds as *least-favorable decisions*. The practicability of the BDF is promoted by the proposed guideline on how to set and fine-tune decision policies formalized by prior and cost beliefs: verbal scales are the basis to finding initial LLR thresholds that are fine-tunable by addressing the mutability of prior and cost ratios. To further interrelate similarities and outline differences of both perspectives, two novel performance plots are proposed, namely BET and NECE plots. BET plots interrelate trade-offs in prior and cost beliefs with trade-offs in error rates, such that vertical and horizontal changes on the BET plot canvas correspond to changes in LLR thresholds whilst both axes depict Type I and Type II error rates. For the conventional error rate based paradigm in performance assessment, BET plots reveal betting odds in the latent decision subspace that is formalized by the BDF. In contrast to BET plots, solely reporting on discrimination performance regarding error rates, NECE plots report on discrimination and calibration performance regarding information gains that result from employing a system to decision making. As the default information provided by a coin tossing system varies depending on the prior probability of each class, as does the empirical cross-entropy of a recognition system. NECE plots visualize the

ratio to default cross-entropy from the empirical information gain in terms of discrimination and calibration performance. As such, one can choose the recognition system that provides the most information to decision making across ranges of prior class probabilities, i.e., across more class \mathcal{A} or more class \mathcal{B} environments.

For classifying performance visualization (established ones and proposed in this work), a taxonomy is introduced. This taxonomy provides an overview on performance visualizations following the perspective of the BDF paradigm. Performance visualizations that are well-established within the fields of speaker recognition, forensic evaluation, and biometric standardization are put into context. Thereby, three performance assessment concepts are identified: the performance reporting type, which is subdivided into *discrimination* and *discrimination and calibration*; the performance reporting scope, which is subdivided into *error rates* and *information gains*; and the targeted audience, which is subdivided into *analytics* and *reporting*.

To the following discourse, the preceding chapter served different purposes. First, when employing Bayesian methods in machine learning of recognition systems, the same principles hold for the evaluation of these recognition systems. Second, gaps are bridged between speaker recognition, forensic evaluation, and biometrics standardization communities. To accommodate the perspectives on performance assessment of these communities, four figures of merit are examined in the following part of this dissertation, primarily C_{llr} , C_{llr}^{\min} , secondarily EER and FMR₁₀₀. Third, the contribution of this dissertation is towards the applicability of LLRs to recognition systems from the definition of decision requirements; from outline, design, implementation, research and development of binary decision systems to the assessment of formalized decision requirements. When facing unconstrained environments, it is very helpful to formalize the performance of speaker recognition systems (to be employed in forensic and commercial application scenarios). As the degrade in the quality of environmental conditions causes performance degrades, these degrades are linkable to specific threshold values (no matter the design of a recognition system). In this dissertation, [quality vectors \(q-vectors\)](#) are proposed to inform on the change between environmental conditions to recalibrate LLR scores and thresholds. This chapter elaborated on the impact of the meaning of this recalibration on both: the *strength of evidence* and the *least-favorable decision* being made.

ON THE INFLUENCE OF DURATION AND NOISE

This chapter addresses the assessment of pre-comparison measures for state-of-the-art speaker recognition systems facing unconstrained acoustic environments. The acoustic feature extraction for a given system is assumed to be fixed; the baseline system follows the [intermediate-sized vector \(i-vector\)](#) paradigm with [probabilistic linear discriminant analysis \(PLDA\)](#) comparison. The following research question is targeted:

Can the impact of environmental conditions in terms of [biometric distinctiveness](#)—as voice references aggregate—and the robustness of voice sample segmentation decisions be quantified before the comparison subsystem?

Different duration and noise conditions are examined regarding their effect to biometric voice samples and to acoustic feature extraction. Thus, this chapter provides context to the experimental study described in the following chapter. Firstly, an estimate to pre-comparison information is proposed, estimating the biometric distinctiveness in terms of entropy. Thereby, the collision probability of [subjects](#) is depicted in the i-vector feature space under different sample durations, i.e., different levels of biometric sample completeness. Secondly, a novel measure for voice segmentation robustness is proposed based on an analysis of noise impacts towards [voice activity detection \(VAD\)](#), i.e., the segmentation of voice samples into *speech* and *non-speech*. Eventually, the scope of the following research is restricted towards optimal VAD in order to keep experimental analyses comparatively fair. In contrast, by including processing artefacts that result from this early signal processing step (the VAD) in analyses at later stages, the argumentation of causal complexity is induced, increasing the reporting complexity of the analyses. As current VAD technology leads to different segmentation decisions by varying quality conditions (as shown in this chapter), these impacts would delude the conclusion drawn from analyses concerning the comparison and decision subsystems. Thus, this chapter concerns the performance degrade observable in the signal processing subsystem to put preceding experiments into context. These experiments solely aim at improving decision making by the normalization and calibration of scores that result from a (fixed) comparison subsystem.

5.1 BIOMETRIC DISTINCTIVENESS OF VOICE SAMPLES

The security or evidence level of an automated recognition system can be depicted in terms of the system's contribution to decision making for forensic or commercial applications, see section 2.4. However, as comparisons differentiate between subjects in the feature space¹⁰⁸, additional information about the underlying feature space can be utilized [211, 212]. Regarding the domain of comparison scores, the work in [115] focusses on **empirical cross-entropy (ECE)** analysis across different recognition systems. In contrast, the biometric information within a given feature space is reported by measuring the relative entropy in the feature space as the divergence between the empirical feature distribution of one subject compared to the generalized feature space of all other subjects [211]. The benefit of this divergence analysis is that it estimates the *collision probability* (when two subjects share the same features). Exemplarily, the collision probability for passwords is derived.

Example: Password Entropy & Collision Probability

The *entropy of passwords* or PINs $H(\text{string})$ can, in comparison, be computed by the string length L and the number N of different string characters that can occur [213]:

$$H(\text{string}) = L \log_2 N. \quad (5.1)$$

Thus, 4-digit PINs have an entropy of 13.3 bits. Compared to more secure passwords with at least 128 bits, users need to remember passwords of $L = 17$ characters (including special characters, assuming all printable extended ASCII codes, such that $N = 224$). Furthermore, when targeting high-evidence and high-secure systems, it is valuable to know when collisions will ideally occur. This information can be derived directly from a feature space's entropy $H(\text{space})$ as the probability p_{col} [42]:

$$p_{\text{col}}(\text{space}) = 2^{-H(\text{space})}, \quad (5.2)$$

where $p_{\text{col}}(\text{PIN}) \approx 1 \times 10^{-4}$, $p_{\text{col}}(\text{password}) \approx 3 \times 10^{-39}$ for 4-digit PINs and secure passwords, respectively.

Compared to passwords, which have no intra-variability as such, biometric systems need to address within-subject variance as well. Passwords require to be remembered, whereas in biometrics, features of the **biometric capture subject** herself are compared to the known and deposited reference features representing biologic and behavioral characteristics [214].

¹⁰⁸ Parts of this section are based on a collaborative work with Rahim Saeidi, Christian Rathgeb and Christoph Busch [69].

The entropy of voice acoustic features (i-vector features) is investigated. As i-vectors represent the characteristic factors of an acoustic subspace based on sufficient statistics, the certainty of the subspace point estimate increases by observing more data, i.e., longer speech durations. In this section, the following hypotheses are investigated:

- The relative entropy across subjects can predict the performance estimate in the front-end i-vector feature space, such that a preliminary to the estimation of discrimination performance can be derived for speaker recognition systems in terms of generalized discrimination information as C_{llr}^{\min} .
- When examining different degradations of voice sample completeness, represented by the duration of a voice sample, the accumulation of biometric probe information cannot only be resembled in the score domain but also in the feature space domain.
- The biometric distinctiveness can be approximated by the cross-entropy between subjects in the *acoustic* feature subspace. This estimate of biometric distinctiveness is independent of the employed biometric feature extraction and comparison. As such, insights into the (posterior) collision probability of subjects on (acoustic) speech data are provided. A figure of merit on the biometric distinctiveness would depict not only the information richness of the biometric characteristic but also the difficulty of a comparison task (when comparing across evaluation datasets or environments), i.e., it would summarize different perspectives on the expected level of security before any *biometric* feature extraction or comparison was conducted.

5.1.1 Feature Space Information

Estimations for biometric information were done, inter alia, by Ratha et al. [42], Daugman [212], and Adler et al. [211]. Adler et al. also refer to the biometric information as a measurement for the biometric distinctiveness. The approaches rely on collision estimations by brute force, estimating the number of independent bits on binarized feature vectors and the relative entropy between genuine and impostor subspaces.

Ratha et al. [42] investigated the probability of guessing features correctly by random feature generation (by brute force). For fingerprints, they evaluate the total number of possible variations for K minutiae locations, m minutiae, and d number of minutiae orientations, thereby formulating the collision probability as $1 / \left(\binom{K}{m} d^m \right)$, from which entropy can be measured using Eq. (5.2). This approach addresses the robustness of a feature space on brute force attacks rather than a feature space's ability to distinguish between subjects.

Daugman [212] analyzed binary iris features in which the Hamming distance is used for comparing all subjects of a database to each other. The author relates the score distribution to a *Bernoulli experiment* with $N = \mu(1 - \mu)/\sigma^2$ degrees-of-freedom, where μ is the observed Hamming distance mean value and σ^2 is the variance, respectively. A feature space's entropy is referred to as N , which represents the amount of coin tosses needed for a feature space collision. This method describes the unique feature space elements of a binary feature space. Relating to it, Adler et al. [211] argue that the question *to what extent biometric characteristics are distinctive* needs to be addressed more, since the distinctiveness is provided by features (from an inference process) and not by the mere number of feature space elements. Notably, the feature space is bound to the acoustic feature extraction. As such, the choice of other acoustic features can yield higher or lower fidelity.¹⁰⁹

Adler et al. [211] introduce a measurement for *biometric information* that addresses the inter-subject information of features \mathbf{x} . The latter is measured by the *Kullback-Leibler divergence* $D(p||q)$ of the intra-subject distribution $p(\mathbf{x})$ and the inter-subject distribution $q(\mathbf{x})$. It represents the needed extra information (in bit) to represent $p(\mathbf{x})$ w.r.t. $q(\mathbf{x})$:

$$D(p||q) = \int_{\mathbf{x}=-\infty}^{\infty} p(\mathbf{x}) \log_2 \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}. \quad (5.3)$$

The p distribution represents a subject's feature subspace, while the q distribution represents the feature space of all other subjects. It is assumed that p and q follow a Gaussian distribution with parameters $p(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ and $q(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$, respectively. By using the Gaussian model, the Kullback-Leibler divergence represents a lower bound to the estimated relative entropy and Eq. (5.3) can be formulated as [211]:

$$D(p||q) = k(\lambda + \text{trace}((\boldsymbol{\Sigma}_p + \mathbf{T}) \boldsymbol{\Sigma}_q^{-1} - \mathcal{I}))$$

with $k = \log_2 \sqrt{e}$, $\lambda = \log \frac{|\boldsymbol{\Sigma}_q|}{|\boldsymbol{\Sigma}_p|}$, $\mathbf{T} = (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^T (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)$. (5.4)

The relative subspace entropy $H(p)$ is computed by the average relative subject entropy. Fig. 5.1 illustrates how the Gaussian model acts as a lower bound compared to a more sophisticated model, i.e., a *Gaussian mixture model (GMM)* over q on Gaussian-distributed exemplary data.

In order to estimate each subject's relative entropy significantly, Adler et al. [211] refer to two regularization approaches:

a) *Regularization for degenerated features*: high-dimensional feature spaces are usually extracted from samples, e.g., with $F = 400$ dimensions, while the analyzed database may only contain a couple

¹⁰⁹ The term *fidelity* is an expression of how accurately a biometric sample represents its source biometric characteristic [215].

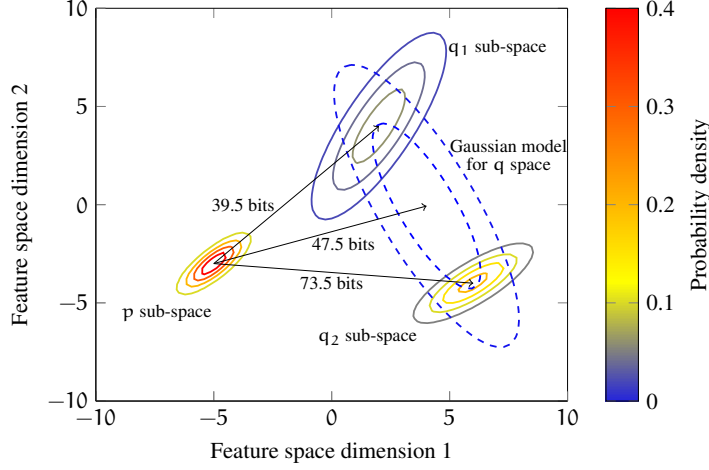


Figure 5.1: Estimating the lower bound of relative entropy by using a generalizing Gaussian model compared to a more detailed GMM on exemplary data. The single-Gaussian model estimates a lower bound of 47.5 bits, while the GMM's estimation is calculated by the mean of each GMM component's relative entropy, i.e., as the average distance of the subspaces of subjects q_1, q_2 to p with 39.5 bits and 73.5 bits, respectively: 56.5 bits.

of samples per subject, e.g., $N_p = 10$. In order to significantly estimate entropy, the feature space is transformed into a G -dimensional space by principal component analysis (PCA), where $G \leq F$. The PCA is performed on the q distribution covariance by singular value decomposition (SVD) since the q distribution is much more accurately estimated than the p distribution, such that:

$$\mathbf{U} \mathbf{S}_q \mathbf{V}^T = \text{svd}(\mathbf{\Sigma}_q). \quad (5.5)$$

The matrices $\mathbf{U}, \mathbf{S}_q, \mathbf{V}$ are truncated to the dimension G . This is done according to an adaptive impact threshold considering the PCA-impact of the first element $10^{-10}[\mathbf{S}_q]_{1,1}$. Elements are truncated at $[\mathbf{S}_q]_{j,j} < 10^{-10}[\mathbf{S}_q]_{1,1}$. Then, the subject's PCA feature space covariance \mathbf{S}_p is computed by:

$$\mathbf{S}_p = \mathbf{U}^T \mathbf{\Sigma}_p \mathbf{V}, \quad (5.6)$$

and the Kullback-Leibler divergence is restated as:

$$D(p||q) = k(\nu + \text{trace}(\mathbf{U}((\mathbf{S}_p + \mathbf{S}_t) \mathbf{S}_q^{-1}) - \mathbf{J}) \mathbf{V}^T) \\ \text{with } \nu = \log \frac{|\mathbf{S}_q|}{|\mathbf{S}_p|} \text{ and } \mathbf{S}_t = \mathbf{U}^T \mathbf{T} \mathbf{V}. \quad (5.7)$$

b) Regularization for insufficient data: given N_p samples, covariance estimations on $G \geq N_p$ will lead to singular $\mathbf{\Sigma}_p$ and will let the entropy diverge to ∞ . In order to avoid ill-disposed $\mathbf{\Sigma}_p$, non-diagonal elements $[\mathbf{\Sigma}_p]_{i,j}$ are set to zero at $i, j \geq N_p$, e.g., on $N_p = 10$ all non-diagonal covariances with column or row indexes $i, j \geq 10$ are zeroed, while the diagonal variances remain.

5.1.2 Contribution: Analysis of Feature Space Information

This regularization scheme needs to be extended since the following analysis requires the (covariance) matrix to be positive finite (which covariance matrices naturally are), but the above regularization can diminish the matrix property. Adler et al. [211] refer to a database on which 16 samples are distributed for each subject, such that covariance estimations are much more sufficient compared to the case of varying sample amounts per subject with $N_p \leq 10$. Thus, the regularization scheme was extended by:

c) *Regularization for ill-conditioned PCA covariances*: non-diagonal elements $[\Sigma_p]_{i,j}$ are iteratively zeroed until Σ_p is positive finite.

d) *Regularization for insufficient sample amount*: mean and covariance estimations need to be estimated from a proper amount of samples, which may vary in databases. Given the properties of most databases, the term *proper* is denoted in this study, such that only subjects with at least $N_p = 10$ samples are examined.

For analytic purposes of estimating the biometric information of state-of-the-art speaker recognition in a duration-sensitive manner, duration-variable p subspaces were compared with full duration q spaces simulating the automatic recognition case, in which full reference *i-vectors* are compared to probe *i-vectors* of all duration groups, see section 3.2.2.1.

Fig. 5.2 and Tab. 5.1 compare the relative entropy among the duration scenarios (full-versus-5/10/20/40/full) and show correlations to the biometric and score cross-entropy performance of a corresponding PLDA comparator with 400 speaker factors. The discrimination performance is reported in terms of the equal error rate (EER), the FNMR at a 1% FMR (FMR₁₀₀), and C_{llr}^{\min} .

Table 5.1: Relative entropy and performance comparison of mixed gender PLDA recognition.

Duration group	Entropy (in bit)				PLDA (400)		
	μ	σ	min	max	EER	FMR ₁₀₀	C_{llr}^{\min}
full-5	127.2	24.0	71.5	226.6	17.0%	66.7%	0.529
full-10	124.3	28.1	65.0	254.8	8.7%	31.6%	0.296
full-20	135.5	35.3	63.2	319.0	4.1%	9.8%	0.147
full-40	155.0	43.1	71.1	421.9	2.1%	3.2%	0.078
full-full	182.1	50.0	88.7	471.6	1.7%	2.1%	0.069

In general, biometric information increases by the speech duration conveyed in the voice sample, which results in better speaker verification performances and lower C_{llr}^{\min} . This behavior is expected as *i-vectors* gain more significance by duration. The standard deviation of the subject-wise relative entropy and the maximum entropy increase. According to the experiments in this work, it is observed that the

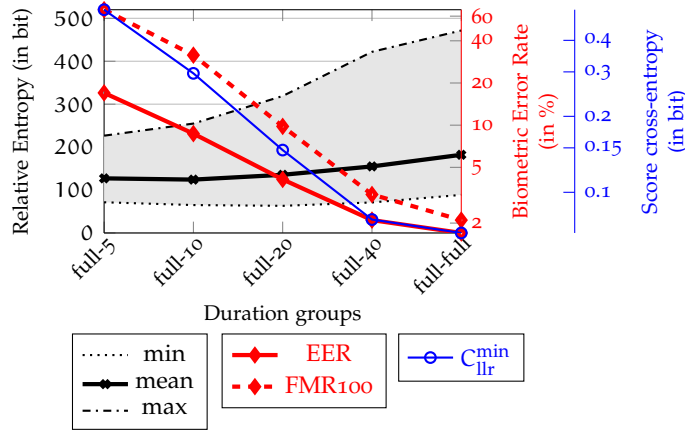


Figure 5.2: Comparison of feature and score domain relative entropy along with speaker recognition performance among the duration groups in common recognition scenario (full versus variable duration) using PLDA with 400 speaker factors on mixed gender.

lowest entropy can be estimated to be as low as 63.2 bits for short duration and 88.7 bits for full segments. The exact numbers of entropy for different system setups could be different, but it is deemed that the trend would be consistent. For face recognition, particularly for a feature fusion, Adler et al. [211] refer to an average of 46.9 bits. The mean of the calculated entropy shows a minimum of 124.3 bits for the full-10 condition—aside from intra-variability. Short speech samples can compete with 128 bits-strong passwords in terms of feature space entropy. The biometric information of full segments yields the highest mean entropy value of 182.1 bits. In gender-dependent analysis, similar results are obtained. The highest relative entropy on the female and male subsets for full segments exceeded 300 bits and 400 bits, respectively.

In order to provide more detailed information about the actual relative entropy of respective subjects, Fig. 5.3 visualizes the duration-based accumulation of relative entropy by each subject. The relative full-full entropy normalizes the relative entropy of all duration conditions. Aside from a few outliers¹¹⁰ that feature more biometric information on shorter samples, the vast majority of all relative entropies is within 50% to 100% of the subject-according full-full entropy. Overall, the subject discrimination in terms of a subject's rel-

¹¹⁰ The more acoustic features are extracted from these outlier subjects, the less distinctive their representation becomes in the light of all other subjects. As acoustic i-vectors resemble the expected offset of acoustic features from a [universal background model \(UBM\)](#), the i-vector representations of these subjects get closer to the representations of other subjects when more speech data is observed. Effectively, acoustic i-vectors are a fixed-length audio representation applicable to many recognition tasks. By projecting acoustic i-vectors to biometric i-vectors, these outlier observations diminish—the baseline systems of preceding experiments therefore employ linear discriminant analysis.

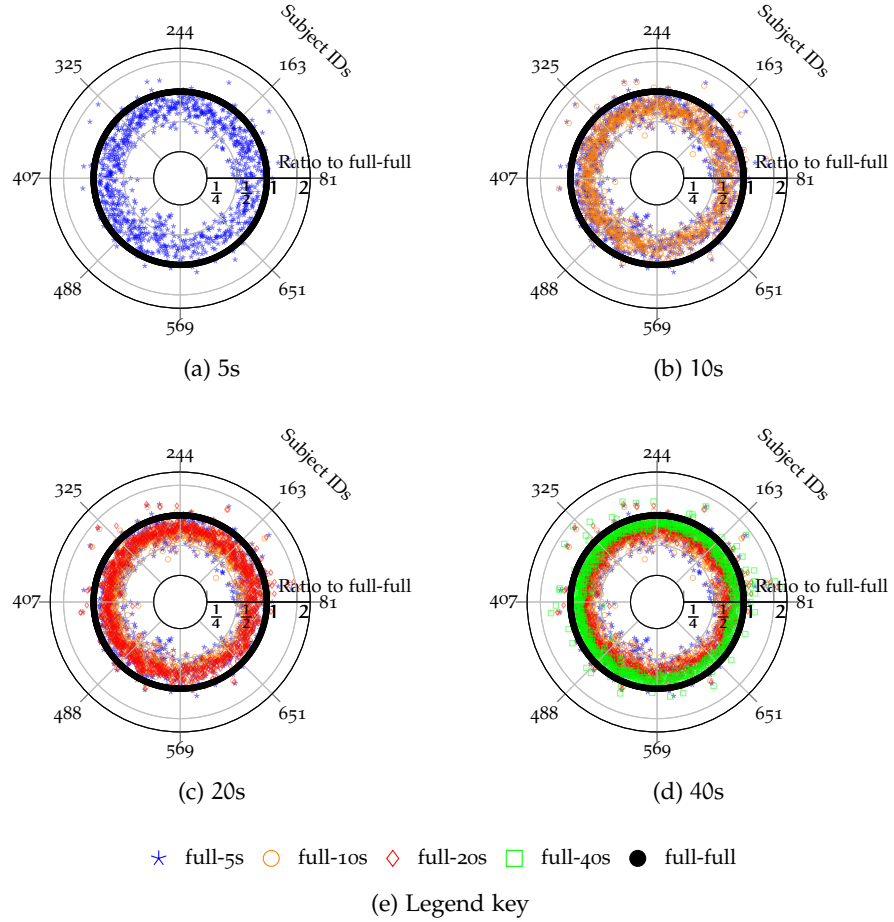


Figure 5.3: Speaker subspace accumulation by duration: relative entropy is normalized subject-wise by the according full-full entropy, the accumulation is logarithmically visualized by ratios, i.e., all full segment ratios are 1, perishing actual entropy value comparisons.

ative entropy accumulates by increasing duration. In comparison to other duration conditions, however, relative entropy of the full-5 condition is more widely distributed and partly reach full-full level. Further, full-10/20/40 relative entropy values accumulate continuously, while there is a gain from full-40 to full-full among the vast majority of all subjects.

In contrast to existing literature on entropy in speaker recognition which merely focuses on the score level, e.g., [36, 115], this work emphasizes on the feature level. It is demonstrated that current speaker recognition feature spaces reach the relative entropy level of 128bits-strong passwords already at 20s of speech, where the recognition performance is acceptable. The generalized collision probability of i-vector based speaker recognition can be estimated as $p_{\text{col}}(\text{voice}_{127.2\text{bits}}) \approx 5 \times 10^{-39}$ for short samples and as $p_{\text{col}}(\text{voice}_{182.1\text{bits}}) \approx 2 \times 10^{-55}$ for long samples. As such, automated speaker recognition is a viable instrument for forensic investigations.

From an industrial perspective, voice is found to be a suitable biometric characteristic for user-friendly high-security commercial authentication mechanisms, such as e-banking. Concerning i-vectors as latent variables which are estimated alongside their uncertainty, one may, as an alternative to the covariance regularization, directly use the precision of the i-vector extraction process. Further gains are expected by fusing i-vectors stemming from different speech signal features.

5.2 ANALYSIS ON THE SEGMENTATION OF VOICE SAMPLES

The segmentation of voice samples into frames of speech and non-speech is referred to as **voice activity detection (VAD)**. Based on VAD-selected acoustic features, biometric features are extracted and compared. The decision robustness of VAD segmentation is fundamental to the performance of a speaker recognition system. Stable segmentation decisions are important to the reliability of biometric systems in unconstrained environments, such as in mobile banking using voice recognition or automated forensic speaker recognition.

In this section¹¹¹, the decision robustness of VAD algorithms is examined targeting mobile telephone data and concerning different noise environments, particularly white, pink, car, and babble noise. White and pink noise represent random channel effects and natural environment backgrounds, respectively. As the noise level is depicted in terms of an **signal-to-noise ratio (SNR)**, i.e., comparatively reflecting the energy levels of speech and noise, conventional energy-based VAD algorithms may be prone to unconstrained acoustic environments. Thereby, the energy accumulates over all frequencies, generalizing the information contained in a signal. In this section, the following hypotheses are investigated:

- Energy based VADs with fixed decision thresholds are prone to low-SNR levels, especially when assuming high-SNR (speech) signals, causing instable VAD decisions, where more robust VAD schemes should benefit from sample adaptive thresholds of rather detailed frequency analyses.
- Preliminary to the estimation of discrimination performance for voice biometric systems, e.g., by the EER, FMR₁₀₀ or C_{llr}^{\min} , a measure for decision robustness of VAD algorithms can aid the selection of VAD algorithms when seeking stable voice sample segmentations and when unconstrained acoustic environments are targeted.

¹¹¹ Parts of this section are based on a collaborative work with Reiner Bamberger and Christoph Busch [68], which emerged from the collaboration with Reiner over his master's thesis [216]. The presented analysis motivates to assume perfect VAD recognition for further research studies of this dissertation, such that depending experimental results can be compared fairly.

5.2.1 Overview: Voice Activity Detection Algorithms

VAD algorithms are introduced in the field of [automatic speech recognition \(ASR\)](#): in order to recognize the verbalized text with less computational effort, speech is segmented into relevant parts, i.e., words by excluding silence, noise, and non-speech sounds. Therefore, VADs are conventionally designed, such that speech is not clipped and non-speech is not falsely segmented as speech.

In this study, emphasis is put on the *ITU recommendation P.56* [217]¹¹², the *Voicebox VAD (VBX)* [218–220], the *simple real-time VAD (SRT)* [221], the *low-complexity variable framerate VAD (VFR)* [222], the *practical, self-adaptive VAD (PSA)* [223], and the *unsupervised GMM-based VAD (USG)* [224], for which an extension is proposed, and the *most dominant frequency component unsupervised GMM (MUG)*, incorporating features from the simple real-time VAD approach.

5.2.1.1 ITU Recommendation P.56

ITU P.56 [217] contains a VAD for telephone speech transmission quality in real-time applications. P.56 is a multi-stage VAD. Firstly, a two-stage exponential averaging on the rectified signal values is performed. Secondly, initial VAD decisions are made by fixed threshold comparison. Thereby, frames are processed in a geometric progression scheme with accumulative activity and hangover counters. The P.56 hangover scheme delays speech to non-speech decisions by 0.2 s, preserving low-energy speech at the end of utterances. Thirdly, activity levels of frames are estimated based on activity counters, which are finally compared to a sample-adaptive threshold derived from the long-term energy and a 15.9 dB margin.

5.2.1.2 Vocebox VAD

The Voicebox VAD [218] is a first-order Markov process modeling of speech with minimum statistic noise estimation. It extends the VAD of [219] by conducting a frame-based [log-likelihood ratio \(LLR\)](#) decision for speech and non-speech propositions examining a-posteriori SNRs estimates based on the power spectrum after discrete Fourier transform (DFT). The noise spectrum is estimated using minimum statistics noise estimation (MSNE) [220] instead of the minimum mean-square error (MMSE) estimator. Thereby, spectral minima are tracked in each frequency band without any speech or non-speech assumptions. The power spectral density (psd) estimation is smoothed by a conditional mean square error estimator. A-priori and a-posteriori SNRs are computed for each frequency band w.r.t. the variance of the smoothed and bias-compensated psd estimation. A

¹¹² P.56 (03/93) is succeeded by P.56 (12/11) with changes in annexes only. This work refers to the P.56 (03/93) Voicebox implementation [218].

Hidden Markov Model (HMM) based hangover scheme is conducted, making decisions for current states also dependent on previous observations. Speech decisions are conducted on a speech posterior probability threshold of 70%. Contrary to conventional hangover schemes, delaying transitions in speech to non-speech decisions, the property of strong correlations in consecutive occurrences of speech frames is modeled explicitly by the VBX VAD [219].

5.2.1.3 Simple Real-Time VAD

Targeting a simple, efficient, and robust algorithm, [221] propose an easy-to-implement and low-complexity VAD for real-time applications based on short term features, i.e., the short-term energy, the spectral flatness measurement (SFM, in dB), and the most dominant frequency component (MDFC), where SFM represents the dB-domain ratio of the geometric mean to the arithmetic mean of the speech spectrum and MDFC represents the frequency corresponding to the maximum spectral value. In [221], thresholds are estimated for each VAD feature based on the minimum feature values within the first 30 frames, assuming them to partially contain non-speech sequences. SRT decides on speech if one of the following votes is positive: short-term energies surpass the minEnergy by an adaptive margin, MDFC surpasses the minMDFC by 185 frequencies, or SFMs surpass the minSFM by 5, i.e., the geometric spectral mean is favored over the arithmetic spectral mean by a factor of $\sqrt{10}$. Energy minima E_{\min} estimates a frame-wise increase by the number of consecutively observed non-speech frames, such that an adaptive threshold is computed as $E_{\text{thres}} = 40 \log(E_{\min})$. SRT examines frames of 10 ms.

5.2.1.4 Low-Complexity Variable Framerate VAD

In contrast to the conventional frame window and hop size setups in speech processing (25 ms and 10 ms) assuming speech signals to have stationary behavior in short time segments, VFR VAD [222] assigns higher frame rates to fast changing and lower frame rates to rather steady events, e.g., consonants versus vowels or silence. Thereby, frames are examined with a frame shift of 100 Hz (1 ms hops), emphasizing on reliable regions in noisy speech. VAD decisions are carried out on a-posteriori SNR estimated distances of consecutive frames: if accumulated distances of non-speech frames surpass an frame-adaptive threshold, frame segments are denoted as speech. VFR preserves sigmoidal turning points between 15 dB and 20 dB. In the online available source code to [222], the VFR VAD decision is outvoted if the posterior probability of a frame being voiced is larger than 25% by utilizing the Voicebox pitch tracker [218].

5.2.1.5 Practical Self-Adaptive VAD

Targeting [speaker recognition evaluations \(SREs\)](#) of the [US National Institute of Standards and Technology \(NIST\)](#), Kinnunen and Rajan [223] propose an unsupervised, self-adaptive, and practical VAD based on [mel-frequency cepstral coefficients \(MFCCs\)](#) ψ_t of frame t and frame energies. Speech signals are denoised and enhanced by spectral subtractions in magnitude domain, power domain, and Wiener filtering utilizing MSNE for noise tracking. MFCCs of the frames associated to 10% of the lowest and highest clean energy values are utilized in order to train two GMMs representing speech λ^{speech} and non-speech $\lambda^{non-speech}$, respectively, which take the form of (cf. Eq. (2.63)):

$$\Pr(\psi_t | \lambda^{speech, non-speech}) = \sum_{c=1}^C w_c \mathcal{N}(\psi_t | \mu_c, \Sigma_c) \quad (5.8)$$

with mixture weights w_c , component means μ_c , and covariances Σ_c , where $\lambda^{speech}, \lambda^{non-speech}$ have the same number of components C for simplicity. GMMs are trained using k-means in order to retain low complexity with $C = 16$ codevectors (components), solely representing 12 static MFCCs (including the zeroth coefficient) without any normalizations nor deltas. VAD decisions are conducted within the [Bayesian decision framework \(BDF\)](#) assuming equal priors and costs, such that the LLR computation reduces to the nearest-neighbor rule, i.e., to a vector quantization based approach:

$$\min_c \|\psi_t - \mu_c^{speech}\|^2 \leq \min_c \|\psi_t - \mu_c^{non-speech}\|^2, \quad (5.9)$$

where a simple energy-based VAD decision $\log E(t) \geq -75$ dB needs to hold as well in order to consider a frame as speech.

5.2.1.6 Unsupervised GMM-based VAD

In [224], an unsupervised GMM based VAD based on a similar design to the PSA VAD is proposed [223], where VAD decision making is conducted by LLR and energy decisions and followed by an finite state machine (FSM) based hangover scheme. Here, *rastamat* [174] is used for computing energy values. The energy decisions are conducted after smoothing $\log E(t)$ values by a 9-frame sliding window moving average filter, where the energy threshold η_E is the average of the values of the 20% and 80% quantiles of sorted $\log E(t)$ values. Similarly, the sample-adaptive LLR threshold η_{LLR} is derived from a 23-frame smoothing. As in the PSA VAD [223], both speech votes are required here for considering a frame as speech. Finally, a hangover scheme is applied in two distinct ways to avoid the loss of speech segments that might be incorrectly labelled preliminary as acoustic

noise: transitions from non-speech to speech states are delayed. In order not to move into the speech state due to false alarms, all frames in the transition phase need to indicate speech, and transitions from speech to non-speech states are delayed, i.e., if noise is indicated, another transition phase prevents speech misses. Motivated by [225], this work refers to 3 and 8 frame states for false alarm and missing VAD transition phases, respectively.

5.2.1.7 MDFC Extension to Unsupervised GMM VAD

GMM-based VADs are motivated by the poor performance of energy-based VADs in low-SNR scenarios, e.g., speech and noise energies are equal on 0 dB SNR, effectively leading to random VAD decisions. Since USG GMMs are self-adaptive to the current speech sample, energy-based selection may result in inadequate-representative training segments, especially in the presence of high-energy noise impulses, such as closing doors or moving nearby objects.¹¹³ Thus, the use of the lowest and highest MDFC instead of energy values is proposed here for initializing speech and non-speech GMMs, respectively. MDFC values are smoothed by a 3-frame sliding window moving average filter before sorting in order to exclude impulsive short-time noises from speech GMM training. The proposed extension is referred to as MDFC-based unsupervised GMM (MUG) VAD.

5.2.2 VAD Metrics in Speech and Speaker Recognition

In speech recognition, VAD metrics represent how much verbalized context is missed in contrast to how many false alarms occur in terms of non-speech that is forwarded to ASR systems. In the beginning of an utterance, a number of speech frames $N_{front-miss}$ could be missed by VAD. During utterances, a number of speech frames $N_{mid-miss}$ could be missed. By contrast, non-speech frames might be mislabeled as speech frames by VAD, leading to false alarms. The number of false alarms occurring directly after an utterance is $N_{over-fa}$, these are *overhanging* speech labels. The number of false alarms in an utterance is N_{fa} . The amounts of correct speech and non-speech decisions are $N_{speech-hits}$, $N_{non-hits}$. The numbers of ground-of-truth frames comprising speech and non-speech are $N_{got-speech}$, $N_{got-non}$.

Conventional VAD metrics [225–228] employ these numbers. Thus, they require frame-wise VAD annotated datasets. These VAD metrics are computed as:

¹¹³ Several samples of the MOBIO database [168] comprise short-time noises at the capture start, which may occur due to, e.g., doors, chairs, or pressing a start recording button.

- the front-end clipping (FEC):

$$\text{FEC} = \frac{N_{\text{front-miss}}}{N_{\text{got-speech}}}, \quad (5.10)$$

- the middle-speech clipping (MSC):

$$\text{MSC} = \frac{N_{\text{mid-miss}}}{N_{\text{got-speech}}}, \quad (5.11)$$

- the non-speech over-hang (OVER):

$$\text{OVER} = \frac{N_{\text{over-fa}}}{N_{\text{got-non}}}, \quad (5.12)$$

- the noise detected as speech (NDS):

$$\text{NDS} = \frac{N_{\text{fa}}}{N_{\text{got-non}}}, \quad (5.13)$$

- the speech, non-speech, and average hit rates (SHR, NHR, AHR):

$$\text{SHR} = \frac{N_{\text{speech-hits}}}{N_{\text{got-speech}}}, \quad (5.14)$$

$$\text{NHR} = \frac{N_{\text{non-hits}}}{N_{\text{got-non}}} = 1 - \text{NDS}, \quad (5.15)$$

$$\text{AHR} = \frac{1}{2} (\text{SHR} + \text{NHR}). \quad (5.16)$$

In speaker recognition, VAD effects are mostly reported regarding their effect to the biometric and decision performance, e.g., in terms of the [EER](#), [FMR₁₀₀](#), [decision cost functions \(DCF_s\)](#), or the goodness of LLRs C_{llr} [223, 224, 229, 230]. Due to GMM and factor analysis based architectures in state-of-the-art speaker recognition [126, 127, 130], contextual information is accumulated, i.e., VAD is as little relevant as possible to reject speech segments in order to estimate sufficient statistics without regard to a segment's context, in which a segment is falsely omitted (missed) or falsely included (false alarms). Thus, FEC, MSC, and OVER are less relevant for the biometric VAD performance assessment. However, these metrics remain useful for developing VADs, such that, exemplarily, FEC and OVER reflect the gains from a two-way hangover scheme, and MSC reflects the benefits of smoothing. Furthermore, AHR equally accounts for SHR and NHR. This is not necessarily optimal from the perspective of retaining speech segments for discriminative biometric recognition, especially if SHR and NHR diverge significantly.

5.2.3 Contribution: Decision Robustness Performance

Targeting VAD performance assessment for unconstrained environments, VAD decisions shall remain stable under changing conditions impacting sample quality, such as varying background noises stemming from different sources. Since VAD decisions are binary, i.e., speech or non-speech, and environmental effects are conventionally examined in certain levels or steps, such as 0 dB, 5 dB, ..., 20 dB and *clean* (when distorting the quality of a clean database), effects on VAD decisions under different environments can be thought of as binary sequences, which have an arbitrary but fixed length for each voice sample as depicted in Fig. 5.4. Given optimal-conditions, e.g., clean and synthetic-distorted samples, each binary VAD decision sequence stemming from distorted signals can be XOR-compared to the clean speech signal and reported in terms of the average Hamming distance \bar{d} depicting the conditional VAD decision error rate for one sample. In order to report VAD decision robustness, the database-average conditional VAD decision error $\mu_{\bar{d}}$ (VDE) is proposed. Other statistic moments, such as variance, skewness, and kurtosis, can aid an \bar{d} -distributional summary and VAD development processes but are not included in further steps for the sake of easier tractability.

Clean	0	0	0	1	1	1	1	0	0	1	1	1	1	0	0		
20 dB	0	0	0	0	1	1	1	0	0	0	1	1	1	1	0	0	$\bar{d} = \frac{1}{16}$
15 dB	0	0	0	0	1	1	1	1	0	0	1	1	0	1	0	0	$\bar{d} = \frac{3}{16}$
10 dB	0	0	0	0	1	1	0	0	0	1	1	1	0	0	0	0	$\bar{d} = \frac{5}{16}$
5 dB	1	1	0	0	1	1	0	0	0	1	1	0	0	1	1	0	$\bar{d} = \frac{8}{16}$
0 dB	1	1	1	0	0	0	0	1	1	1	0	0	0	0	1	1	$\bar{d} = \frac{16}{16}$
Condition	VAD decision examples on 16 segments																

Figure 5.4: VAD decision example under changing environmental conditions with segment-wise VAD votes as *speech* (1) and *non-speech* (0), where *clean* denotes the original sample of good quality.

Experiments are carried out on the 2013 MOBIO SRE task [37], see section 3.2.1. Analyses are conducted regarding the VAD decision robustness in noisy conditions of different SNR levels and the coherence of VAD metrics in terms of sensitivity to the evaluation criteria. The performance of speaker recognition systems using VADs is compared in Tab. 5.2. Here, discrimination performance is reported in terms of the EER, the FMR₁₀₀, and C_{llr}^{\min} . VFR outperforms the other algorithms in EER and C_{llr}^{\min} , while the proposed MUG VAD yields a better FMR₁₀₀ performance. The performance gain of systems to no VAD segmentation applied is moderate since the MOBIO task comprises rather prompted speech instead of phone calls, i.e., samples are pre-segmented due to the prompted scenario.

Table 5.2: VAD algorithm performance comparison to no VAD applied by EER, FMR₁₀₀, and C_{llr}^{\min} on male speaker subset of the MOBIO dev-set on clean condition.

	VAD	P.56	VBX	SRT	VFR	PSA	USG	MUG	no VAD
EER (in %)		11.0	10.9	12.2	10.2	10.9	11.0	10.7	11.9
FMR ₁₀₀ (in %)		42.4	41.1	45.9	40.0	43.7	41.7	39.6	46.7
C_{llr}^{\min}		0.377	0.376	0.376	0.355	0.373	0.377	0.361	0.407

5.2.3.1 VAD Decision Robustness

In order to analyze the impact of noise conditions (source types and SNR levels) to VAD and biometric recognition performance, pink, white, babble, and street noise are examined in 0 dB, 5 dB, ..., 20 dB SNR levels using the Matlab implementations of [218, 231, 232]. Pink noise is described to be ubiquitous in many biological and physical systems [233]. White (Gaussian) noise represents random signals. Babble noise is conducted utilizing all speakers of the MOBIO background set with random sample selection. Street noise stems from the QUT-NOISE-TIMIT corpus [234], which is explicitly designed for the purpose of evaluating VAD performance.

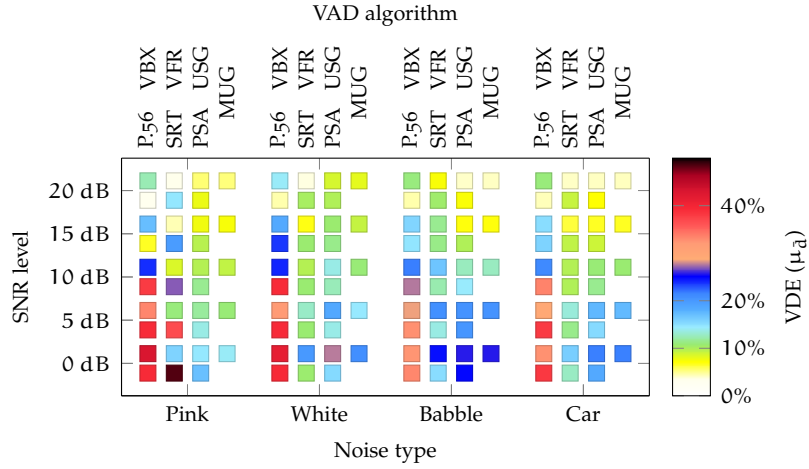


Figure 5.5: VAD decision robustness carried out under noisy samples compared to clean samples on dev-set by the proposed VDE metric μ_d .

Fig. 5.5 depicts the robustness of VAD algorithms regarding the proposed VDE metric, representing the average rate of misconducted VAD segmentation votes. For the majority of VAD algorithms, speech decisions on 15 dB and 20 dB conditions are similarly to the clean condition. On 0 dB and 5 dB, SRT yields the most stable decisions on white, babble, and car noise, and MUG yields the most stable decisions on pink noise. Regarding 10 dB to 20 dB, VFR outperforms other

VADs on pink, white, and street noises, while for babble noise, USG and SRT yield more stable VAD decisions on high-SNRs and 10 dB, respectively. Examining SRE-related VADs, the proposed MUG VAD outperforms PSA and USG on pink, white, and street noise in 0 dB to 15 dB conditions, and on 20 dB pink and white noise. In other conditions, USG outperformed PSA. On condition-averaged VDE, MUG yields 0.120, USG 0.129, and PSA 0.130, whereas VFR and SRT yield 0.113 and 0.157, respectively.

5.2.3.2 Sensitivity Coherence: VAD to Biometric Recognition Metrics

Sensitivity analyses are conducted in order to provide insights into coherence of the proposed VDE metric to biometric and forensic performance. For tractability purposes, noise conditions are pooled by SNR level and the SNR of clean samples is assumed to be 25 dB. Fig. 5.6 depicts the SNR sensitivity of the SRT, VFR, USG, and MUG algorithms. SRT achieves low sensitivity regarding EER and FMR₁₀₀, even though SRT yields the highest EER and FMR₁₀₀ results among all examined VADs, cf. Tab. 5.2. In terms of the proposed VDE metric and FMR₁₀₀, VFR, USG, and MUG VADs perform similarly. Regarding C_{llr}^{\min} and EER, however, USG and MUG result in a more stable performance than VFR, especially in the low-SNR region with average EER sensitivity of 0.76% and 0.80% in EER per 1 dB SNR.

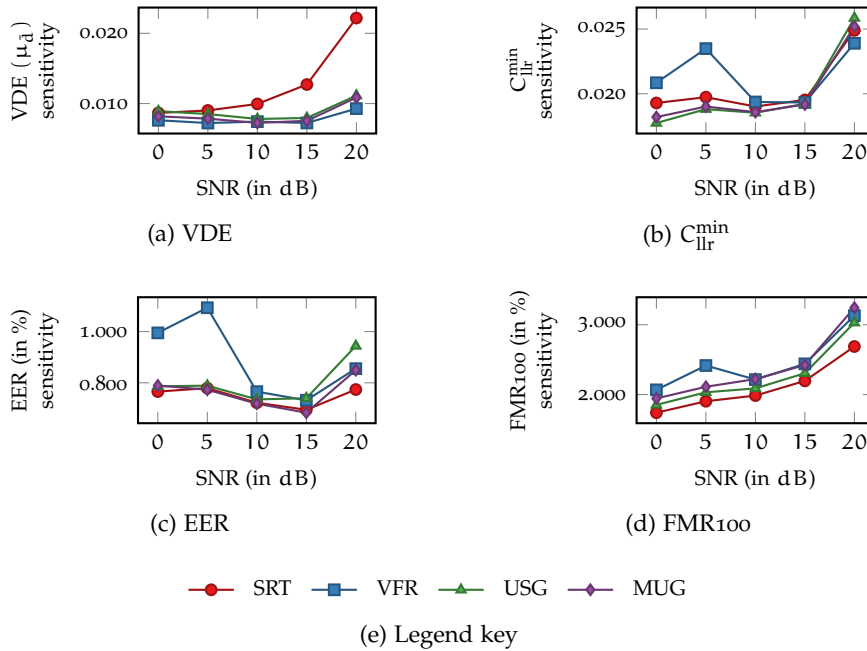


Figure 5.6: Sensitivity of VAD performances to different SNR levels of SRT, VFR, USG, and MUG approaches by (a) VDE, (b) C_{llr}^{\min} , (c) EER, (d) FMR₁₀₀.

VADs are designed for certain applications and target specific environmental constraints, such that none of the examined algorithms is able to outperform other approaches in each analysis. NIST SRE motivated VADs yield more stable segmentation decisions in high-SNR conditions than conventional VADs. However, examining low-SNR conditions, conventional and SRE VADs achieve good performance, particularly the VFR, PSA, USG, and MUG algorithms. For the VFR approach, incoherent sensitivity bumps are observed at 5 dB. This can be attributed to its variable frame rate segmentation—under the presence of noise, ideally shorter frames for plosives might be longer as intended, considering short frame lengths for high-quality speech data (an *incoherence*). Regarding the sensitivity of evaluation criteria, the USG and MUG algorithms yield the least SNR-sensitive results, which is coherent to the conducted VAD decision sensitivity analysis. Moreover, the proposed MUG outperforms other SRE VADs by utilizing beneficial MDPC-features from conventional VADs. The SRT and VFR algorithms partially achieve gains by employing SFM, a-posteriori SNR, and pitch features as features—especially in white, babble, and car noise conditions with low-SNR.

The proposed VDE metric reveals the stability of VAD segmentation decisions under different noise conditions. In contrast to well-established metrics in ASR, the proposed metric examines the average amount of inconsistent VAD decisions on changing environmental conditions, emphasizing on *where* speech frames are falsely recognized. By conducting the proposed analyses approach for examining the decision robustness and evaluation criteria sensitivity of VADs, coherent decisions can be made regarding the applicability of VAD segmentation algorithms to speaker recognition tasks. The proposed metric has limitations regarding the location of false segmentation decisions, which can be examined by conventional VAD metrics. However, decision robustness is more valuable to state-of-the-art speaker recognition methods, in which speech frame statistics are accumulated, i.e., the location of VAD errors remain without impact, whereas unstable VAD decisions lead to different frame samplings forwarded to acoustic and biometric feature extraction. Contrary to conventional VAD metrics, frame-wise annotation voice samples are not required in order to measure VAD performance. Furthermore, the proposed MUG-extension of the USG approach yields promising gains, which are expected to be more extended by incorporating SFM, a-posteriori SNR, and pitch features into the VAD decision process.

5.3 SUMMARY

In this chapter, systems of fixed acoustic feature extraction were examined in unconstrained environments, exemplarily depicting impacts of duration or noise to the biometric distinctiveness and the segmentation of voice samples.

Examining the biometric distinctiveness, a pre-comparison performance measure was introduced to speaker recognition: the relative entropy between subjects provided a bound-estimate to the subject collision probability in the *i-vector* feature space. Analyzing the biometric information among different voice sample durations, the accumulation of voice references was visualized regarding the divergence between subjects rather than by the sample completeness itself. Insights into the difficulty of the biometric comparison task were therefore gained before an evaluation of the actual recognition performance was conducted. Likewise, different acoustic features could be compared regarding the provided biometric information as well as the comparability among features across different biometric modalities and to other authentication features. The correlation between relative entropy of the feature space and the cross-entropy comparison performance proofed the soundness of the proposed figure of merit.

Investigating the segmentation of voice samples, the VAD decision robustness was proposed and another pre-comparison performance measure to speaker recognition provided. Traditional energy-based VADs select audio frames randomly on low-SNR, whereas sample adaptive VADs examining more detailed spectral characteristics provide higher reliability. In order to sustain rather reliable VADs in terms of predictability, the analysis of the sensitivity of performance criteria is proposed—the sensitivity of the proposed decision robustness measure as well as of the sensitivity of, e.g., C_{llr}^{\min} . Even by selecting VAD methods of coherent sensitivity in their performance changes, their decision robustness degrades to the extent that further studies in this dissertation assume oracle (perfect prediction) VAD, leaving research towards robust VADs for future work.

ENHANCING DECISION INFORMATION

This chapter depicts the research progress on unconstrained acoustic environments, targeting the following research question:

Can **quality** mismatches be estimated at pre-comparison stages and aid the Bayesian identity inference in discrimination and calibration performance?

Mutual duration and noise conditions are therefore examined on the assumption of oracle (perfect) **voice activity detection (VAD)** segmentation: no matter the signal distortion, *speech/non-speech* segmentation decisions are ideal. The baseline system used in this chapter is a state-of-the-art speaker recognition system extracting standard **intermediate-sized vectors (i-vectors)** (model representations of the acoustic and the biometric features), which are compared by **probabilistic linear discriminant analysis (PLDA)**, see section 2.5. PLDA is a generative model that is not only capable of discriminating i-vector pairs regarding their **log-likelihood ratio (LLR)** *strength of evidence*, but also of generating i-vector pairs of desired LLRs—trained to infer biometric identities by decomposing i-vectors into biometric and non-biometric components. Thereby, within-class and between-class variability is estimated in the latent subspace: evidence is reported in the form of LLRs based on the inferred similarity and dissimilarity likelihoods rather than reported as scores directly computed from observed features.

The system is trained in a condition pooled fashion, see chapter 3. The duration and **signal-to-noise ratio (SNR)** conditions are thus derived by truncation and synthetic noise degradation.¹¹⁴ The duration and SNR conditions are sampled in a log-linear fashion, motivated by performance loss observations in related work [32, 167]. In contrast to the systems of [32, 167], the baseline system of this dissertation employs a more noise robust extraction of **mel-frequency cepstral coefficient (MFCC)** features by using *maximum-likelihood short-time spectral amplitude* [193] (noise suppression filtering) and *stabilized weighted linear prediction* [192] (robust spectrum estimation). For the sake of easier tractability, duration and noise levels are jointly depicted as *quality* aspects. Here, duration resembles a proxy measure to the completeness of voice samples (the more sufficient statistics are accumulated, the better the discrimination performance) and noise

¹¹⁴ The *Lombard effect* (in order to enhance audibility, speakers increase their vocal effort in the presence of noise) and other speech signal degradation, such as reverberation, are not assessed as this would require real world data, that is not feasible in the experimental setup.

types (AC/CROWD for non-biometric/biometric noise). SNR levels resemble proxy measures to the signal disturbance.

Resulting scores are well-calibrated (LLRs) if the PLDA model assumptions hold. As PLDA is not trained with formalized prior quality probabilities, the proportion of conditions during training needs to predict the condition proportion during evaluation. By narrowing experimental analyses to single quality conditions, these assumptions are violated¹¹⁵, but the requirement of speaker recognition systems to report evidence for good decision making persists. In this dissertation, the biometric system is assumed to be fixed in terms of acoustic and biometric feature extraction as well as biometric comparison. Neither the i-vector nor the PLDA comparator are changeable by a **biometric system owner**, **operator** and **provider** (solely **vendors** can adapt subsystems).

To retain discrimination performance, this chapter proposes a novel score normalization method, adaptive to quality. To retain calibration performance, it proposes a novel score calibration method, also adaptive to quality. For both, this chapter proposes **quality vectors (q-vectors)** that are the basis to either approach. Finally, for the purpose of investigating on the gains and limitations of these q-vectors, analyses are conducted employing deep learning to get insights into the possible gains and limitations of the proposed q-vectors in score normalization.

6.1 SCORE NORMALIZATION: QUALITY ADAPTIVE THRESHOLDS

Score normalization is an effective tool that accounts for the mismatch of reference and probe conditions [163, 164, 235]. Targeting unconstrained acoustic environments, the performance of speaker recognition systems degrades under noisier and shorter duration conditions (Fig. 3.1 and chapter 5), among other signal effects. When dealing with real-life conditions, where the quality of audio recordings in test phase does not match **enrolment** utterance(s) of speakers, a vast broadness of environmental constraints needs to be covered.

Related work emphasizes transfer learning approaches, such as inter-dataset compensation [236] and training comparators in a joint condition (condition pooled) fashion [194, 237]. Acoustic and biometric feature extractors are trained on rather optimal conditions and are then put to recognition tasks on out-domain data. Lately, the robustness of state-of-the-art speaker recognition systems has been addressed regarding feature extractors [238–241], uncertainty-aware

¹¹⁵ As illustrated in Fig. 3.1 for high quality conditions, discrimination performance is high and the calibration loss is neglectable (despite appearing relatively large in the log-log performance plot). For poorer quality conditions, however, discrimination and calibration losses decrease dramatically: the i-vector representation precision shrinks due to the insufficient accumulation (information aggregation and estimation) of the underlying sufficient statistics.

comparators [137, 151] as well as score normalization and calibration schemes employing quality metrics [38, 70, 166, 242].

In this study¹¹⁶, emphasis is put on the quality adaptive score normalization of systems trained in a condition pooled fashion. Rather than targeting a limited number of *constrained* conditions, such as the *unified audio characterization (UAC)* approach in [38], this work targets a host of *unconstrained* conditions. Motivated by the UAC approach [38], q-vectors are proposed in this dissertation: quality estimates are predicted by analyzing (non-biometric) changes in acoustic i-vector features. These quality estimates inform on the posterior probability of each condition. As comparators are trained in a condition pooled fashion, quality estimates are intended to adaptively re-bias scores (scores that are LLRs in pooled quality conditions but are uncalibrated in single conditions as the comparator is uninformed about present conditions) to the right extent to improve the performance in particular conditions.

In this dissertation, *quality* is not interpreted as a *measure of biometric utility* but as *distinguishing (acoustic) characteristics of speech*. For the former, *low quality* implies *limited biometric information*. In this case, all LLRs should reduce towards zero (evidence). For the latter, *distinguishing characteristics* are, e.g., the duration of speech, noise types afflicting speech, and the SNR level of that noise. *Low quality* implies low values of, e.g., duration and SNR.

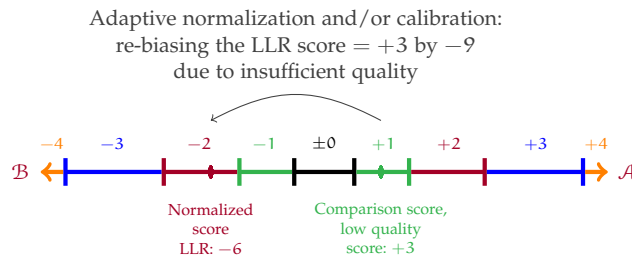


Figure 6.1: Concept: re-biasing thresholds depending on the quality of samples involved in comparison; the illustration also depicts the verbal scale of conclusion, cf. section 4.2: if a sample's quality is insufficient, the related comparison score's value is adjusted depending on the score normalization model.

Score normalization is placed at the transition from comparison to decision subsystems (sections 2.2, 2.4 and 2.5). There, the biometric information on the reference–probe comparison (the evidence) is aggregated to a scalar representation, which is changeable at the end of the comparison subsystem as well as at the start of the decision subsystem. As the biometric evidence is summarized by a score, the comprised biometric information cannot be further enhanced by employing quality estimates. By using quality estimates, however, the in-

¹¹⁶ Parts of this section are based on a collaborative work with Rahim Saeidi, Christian Rathgeb and Christoph Busch [70].

formation divergence between operational and training data is quantifiable, such that scores are jointly normalizable for each condition. The underlying idea is exemplarily illustrated in Fig. 6.1: a score that favors proposition \mathcal{A} is normalized to a score favoring proposition \mathcal{B} (to an LLR), adjusting the evidence reported by a score. Before calibration, this score might have been trained as an LLR but is badly calibrated in the light of condition quality assumptions, diverging between the assessment of a probe and the training of the signal processing and comparator subsystems. To compensate this gap, the comparison subsystem is adapted based on quality estimates. Alternatively to this perspective (on augmenting the comparison subsystem), one might argue that thresholds are adapted at the decision subsystem, depending on each comparison trial: prior and cost beliefs are adjusted between decision policies depending on the quality divergence underlying these policies.

The investigation outlined in this chapter has two goals: to serve as an example on how to examine the performance of formalized decision making of systems operating in the light of unconstrained environments and as a study on the biometric modality *voice*. Therefore, this research focuses solely on duration (sample completeness) and on two noise types, i.e., noise stemming from air conditioning systems (AC) mimicking office noise and noise stemming from multiple speakers in the background, conventionally referred to as babble or crowd noise (CROWD). As depicted in section 3.2.2.2, noise is added, which is why the Lombard effect (among other effects) is not addressed in this work. Additional data utilized in score normalization, i.e., the cohort data (section 2.5), is shortened and degraded by noise as well, resulting in multiple sets of the original cohort data, (section 3.2.2.2). For each (cohort) i-vector, a q-vector is assigned.

In this section, three hypotheses are investigated: i) motivated by the UAC approach, q-vectors should be capable of significantly classifying quality conditions; ii) q-vectors are beneficial for correlating acoustic features, i.e., i-vectors across different quality conditions; and iii) the discrimination performance of conventional score normalization techniques is expandable by employing information about the q-vector similarity.

6.1.1 Unified Audio Characterization Motivated Quality Vectors

In order to establish an automated mechanism for estimating reliable audio quality metrics, a probabilistic scheme based on the UAC approach [38] is proposed to yield posterior probabilities for each condition. For the purpose of estimating condition posteriors, single multivariate Gaussian models $\Lambda_c \sim \mathcal{N}(\mu_c, \Sigma)$, $c = 1, \dots, 55$ are trained in acoustic i-vector space. The models have condition-dependent mean vectors μ_c and share a full covariance matrix Σ . Class-dependent

means are estimated using i-vectors from respective quality conditions and Σ is estimated by pooling all the i-vectors. The resulting vector of condition posterior probabilities¹¹⁷ for an i-vector \mathbf{i} is a q-vector \mathbf{q} , and its elements $q(c)$ are the following posteriors:

$$q(c) = \Pr(\Lambda_c | \mathbf{i}) = \frac{\Pr(\mathbf{i} | \Lambda_c)}{\sum_{c=1}^{55} \Pr(\mathbf{i} | \Lambda_c)}. \quad (6.1)$$

All voice sample representations (references, probes, cohorts) are extended to a pair of an i-vector and a corresponding q-vector.

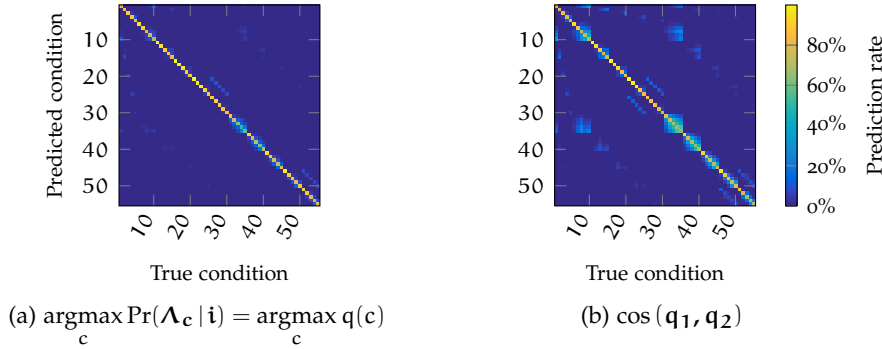


Figure 6.2: Condition confusion matrix on q-vectors for two condition classifiers—likewise *condition linkers* when taking the perspective of the opposing recognition task to classification.

Fig. 6.2 depicts two confusion matrices among all conditions. Tab. 3.4 defines the condition indices. Conditions 31 to 40 comprise the highest signal degradation in terms of SNR and the CROWD noise type and the shortest durations with 5 s and 10 s. These conditions are more likely to be confused with other ones as they relate from the perspective of the signal processing subsystem. For the maximum posterior condition classifier, misclassification rates up to 51% are observed. The vast majority of conditions are far more well-classified, i.e., with misclassification rates less than 20% and correct classification rates up to 99.6%. On AC and CROWD noises, only 10% of 40 s/noisy conditions are recognized as their full/noisy condition equivalents have similar SNR levels. By contrast, a cosine distance similarity classifier operates symmetrically and provides more cross-condition similarities, i.e., more condition misclassifications, which is good for linking conditions with another despite being non-ideal for condition classification. Using the cosine similarity, q-vectors of different conditions are linkable since the cosine operator is a weaker condition classifier than the maximum posterior.

¹¹⁷ Equal prior probabilities are assumed. If the likelihoods of each environmental condition to occur are known, one may introduce different priors for each condition.

6.1.2 Analysis: *i*-vector Pool Mean Shift

In order to measure *i*-vector property changes by signal degradation, *i*-vector condition mean values are examined, raising the question whether cross-condition *i*-vectors share the same mean value or not. In other words, are *i*-vectors of different conditions stemming from the same population, such that they should be treated as samplings from the same condition cluster, or not? Contrary to previous work [235], where *i*-vectors are tested element-wise for shared mean dimensions by the Student *t*-test, this work considers *i*-vector space mean values among different conditions. Therefore, the generalized, multivariate Student *t*-test is used, namely the Hotelling's *T*-squared statistic [243, 244]. In the according statistical test for population-independent means, the null hypothesis states that *i*-vectors share the same mean value among conditions, whereas the alternative hypothesis states that *i*-vector mean values differ amongst conditions. Following the *q*-vector modeling, equal covariances are assumed. The test value of the generalized Student's *t*-test t^2 uses the averaged scatter of both populations \mathbf{W} and is defined as:

$$t^2 = \frac{n_x n_y}{n_x + n_y} (\bar{\mathbf{x}} - \bar{\mathbf{y}})^T \mathbf{W}^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{y}})$$

with

$$\mathbf{W} = \frac{\sum_{i=1}^{n_x} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T + \sum_{i=1}^{n_y} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T}{n_x + n_y - 2},$$

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{n_x} \mathbf{x}_i}{n_x}, \quad \bar{\mathbf{y}} = \frac{\sum_{i=1}^{n_y} \mathbf{y}_i}{n_y}, \quad (6.2)$$

where n_x, n_y are the number of observations on *D*-multivariate data sets \mathbf{x} and \mathbf{y} , respectively. For examining two conditions, the terms \mathbf{x} and \mathbf{y} resemble the acoustic *i*-vectors extracted from each condition. In this experimental setup, *D* equals 200. *P*-values are estimated by the cumulative distribution function *F* of χ^2 distributions [243, 244]:

$$t^2 \sim \chi_D^2, \quad \text{s.t.:} \quad p = 1 - F_{\chi_D^2}(t^2). \quad (6.3)$$

Fig. 6.3 illustrates observed test values between all 55 conditions; *p*-values will result either as exactly one on χ^2 scores of zero or as *p*-values lower than $10^{-13} \approx 0$ indicating high significance. Only same-condition tests result in zero χ^2 scores. Hence, all cross-condition mean shifts are highly significant. Therefore, the *q*-vector elements are indicated to be statistically independent. The underlying model assumes a full, shared, pooled covariance nevertheless. Thus, *q*-vector similarities are expressible in terms of probabilistic divergence and Euclidean distance, e.g., employing the cosine distance similarity.

In other words, the proposed *q*-vectors are capable of classifying sample quality conditions.¹¹⁸ However, the utilization of condition

¹¹⁸ As such, *q*-vectors might be beneficial for the ISO/IEC project family 29794 on *biometric sample quality*. Particularly, a quality score for voice data could be derived (a

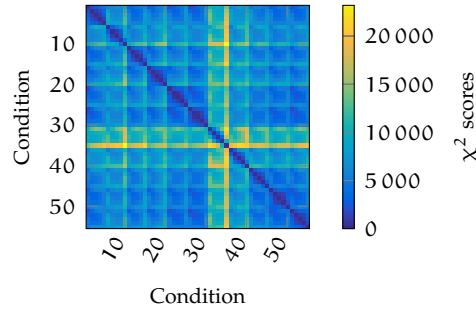


Figure 6.3: Multi-variate Student's t-test: i-vector mean value similarity among conditions.

classification for the purpose of selecting rather suitable subsystems introduces *binning errors* [25], which increase the complexity of system evaluation but mainly decrease the information capacity provided by the original evaluation database to a mere subset. The more conditions are targeted, the lower the average expected information capacity for each condition, i.e., by binning evaluation data, performance estimation is expected to be less accurate. In this study, q-vectors are utilized as quality estimates for score normalization and calibration purposes in a holistic fashion.

6.1.3 Contribution: Unconstrained Cohort based Score Normalization

Conventionally, score normalization in speaker recognition examines the distribution of comparison independent cohort data in order to project a comparison score into a zero-mean and unit variance domain. For the sake of robustness and in order to estimate normalization parameters from a non-skewed Gaussian distribution without tails, the top-n cohort scores are selected.

Contrary to the conventional AS-norm, cf. section 2.5.6, a quality based cohort pre-selection is encouraged: sample q-vectors are derived from condition posterior probabilities. Probe-alike cohort representations are determined by the minimum (symmetric) [Kullback-Leibler divergence](#). It is hence possible to approximate condition-matched cohort sets. Thereby, the theoretical framework on using quality measures [245] is extended by the score normalization stage. Condition-matching cohort selection schemes are expected to not only normalize false matches on references and false non-matches on probes, but also to encounter condition-depending signal degradation.

While conditions are classifiable by the maximum posterior probability, the cohort selection requires a similarity metric to find *nearby* cohort voice samples in the form of q-vectors. Inspired by [246],

potential part 13 of the 29794 family). More signal aspects than duration and SNR need to be investigated.

the symmetric Kullback-Leibler divergence D_{symKL} of two q -vectors q_a, q_b is proposed here for pre-selecting cohorts:

$$D_{\text{symKL}}(q_a || q_b) = \frac{1}{2} \sum_{c=1}^{55} q_a(c) \log \frac{q_a(c)}{q_b(c)} + q_b(c) \log \frac{q_b(c)}{q_a(c)}. \quad (6.4)$$

The closest top- k cohort q -vectors are selected by $\min D_{\text{symKL}}$.

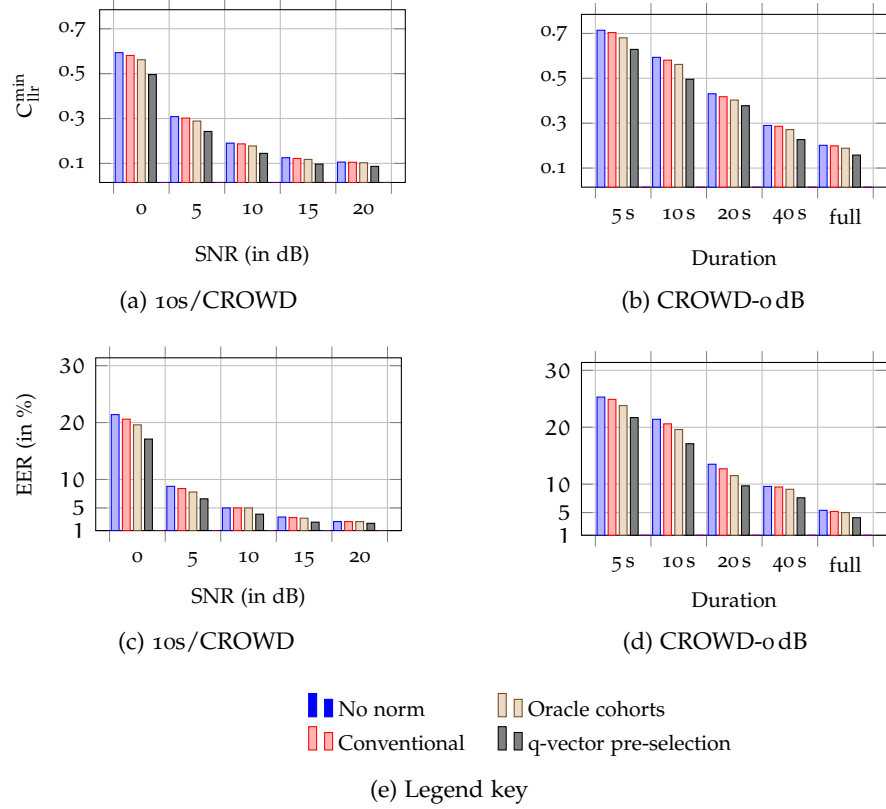


Figure 6.4: C_{llr}^{\min} and EER comparison of conventional AS-norm to oracle cohorts and the proposed pre-selection by q -vectors in extreme conditions.

Aiming at mutual high-degradation conditions, Fig. 6.4 compares AS-norms by SNR levels on 10s/CROWD and by duration groups on CROWD-o dB. The proposed cohort selection significantly outperforms all other systems in C_{llr}^{\min} and equal error rate (EER), including the oracle cohort selection of earlier work [235], proving the soundness of quality based cohort selection. Taking information of other comparisons into account, AS-norm can compensate subject and condition-dependent variances on score domain. Examining C_{llr} performance, the unconstrained AS-norm outperforms the baseline (and the conventional AS-norm) among the vast majority of conditions with relative gains up to 8.2% in C_{llr}^{\min} , 15.9% in EER, and 23.4% in FNMR at a 1% FMR (FMR₁₀₀). The cohort size in terms of top- n selection size, however, has no impact on these metrics, making a

cohort size of 50 interesting for least-effort concerns. A cohort selection scheme is examined seeking reference-alike cohorts \mathfrak{R} and probe-alike cohorts \mathfrak{P} . However, no sufficient gains to the conventional AS-norm are observed, confirming the proposed unconstrained AS-norm approach.

Fig. 6.5 illustrates which conditions and cohort subjects are considered in pre-selection: cohorts having similar noise source, duration, and SNR level are favored, while the vast majority of other conditions is not considered even in a single cohort speaker. The most cohort representations are selected from conditions 36 to 38 (10 s/CROWD-0 dB to 10 dB) and from conditions 39 and 40, completing the block of SNR levels in 10 s/CROWD conditions. Noise source impacts reveal selections representing 10 s/clean and 10 s/AC-0 dB, respectively, from conditions 2 and 11. Duration impacts reveal from selections of conditions 31 to 35, denoting 5 s/CROWD noise conditions. This pattern is also observed on increasing duration, where much more cohort speakers are considered among duration and noise similar conditions by longer probe durations. q-vector based pre-selection of cohort i-vectors provides a more natural perspective to the relevant cohort than the oracle condition or a scalar summary.

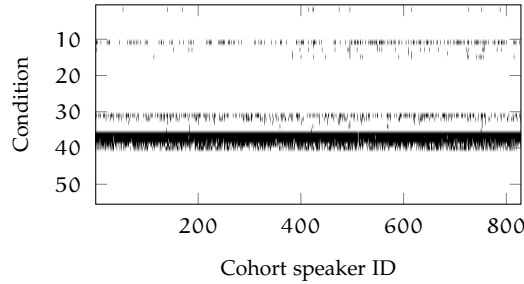


Figure 6.5: Pre-selected cohort subjects and conditions on 10 s/CROWD-0 dB (condition 36) by unique selection (black).

Mutual duration and noise effects severely affect speaker recognition in terms of sample completeness and quality. The condition-informed (unconstrained) AS-norm robustly improves biometric and forensic performances, but it is clearly not capable of reaching the performance on full/clean samples. However, by quality-based cohort pre-selection instead of relying on oracle cohort sets, significant gains in biometric and forensic performance are yielded. Therefore, this approach seems also promising for similar issues, such as domain shift compensation.

6.2 SCORE CALIBRATION USING QUALITY ESTIMATES

Concerning calibration, this section¹¹⁹ solely aims at the calibration of a conventional i-vector system. As the investigation on direct impacts of unconstrained environments to calibration is emphasized, the systems comprising score normalization approaches as well are left for future work. Targeting commercial and forensic application scenarios, robust handling of real world data is still a challenging topic: while commercial applications can focus on known environments and permit sample re-acquisitions from the biometric data subject, the environmental setup in forensic scenarios changes case by case and sample recaptures are not always possible. Furthermore, commercial applications also need to cope with varying conditions due to the rising demand for highly mobile applications facing unconstrained environmental conditions regarding sample acquisition processes. Therefore, a consistent alignment with the [Bayesian decision framework \(BDF\)](#) is fundamental in order to sustain meaningful thresholds—a score calibration yielding *reliable* LLR scores.

Variations in signal quality, i.e., in the probe sample condition, result in different score distributions per condition [166, 167]. While systems are usually calibrated for known scenarios and in fixed condition environments, handling unconstrained conditions imposes well-calibrated decision thresholds among known and unseen conditions.

Thereby, lower miscalibration costs C_{llr}^{mc} are sought compared to conventional calibration, which trains calibration functions on data stemming from an optimal condition (long duration, noise-free). This mismatched/conventional calibration scheme is expected to model low-SNR and short-duration conditions insufficiently and thus to state the lower performance bound in the robustness analysis of this work. By contrast, *oracle* calibration is considered optimal calibration since calibration parameters are trained depending on each condition respectively, requiring calibration functions to be trained for each condition. Hence, the complexity in terms of degrees-of-freedom increases on oracle-condition calibration, when more conditions are considered and further, unseen conditions cannot be calibrated well. Therefore, [sample quality](#) based calibration promises an adequate trade-off between model complexity and accurate approximation of oracle-condition calibration in scenarios facing a wide range of combined noise and duration conditions.

In this section, two hypotheses are investigated: i) score calibration schemes can be enhanced by employing information about the similarity of q-vectors as an alternative to quality measures (regarding calibration gains in C_{llr}^{mc}); and ii) robustness is pertained by q-vector based calibration in comparison to existing quality calibration

¹¹⁹ Parts of this section are based on the collaborative work with Rahim Saeidi, Christian Rathgeb and Christoph Busch [71].

schemes, especially when tough conditions are unknown to calibration training.

6.2.1 Quality Measure Functions (QMFs)

In scenarios targeting different conditions, the conventional linear calibration is observed to be prone to miscalibration when dealing with recognition scores originating from low-quality probe segments [166, 167]. Training calibration parameters condition-dependently leads to inconvenient effects, such as higher system complexity in terms of parameters to train. **Quality measure functions (QMFs)** [32, 166] are introduced in order to account for the quality of reference and probe samples in the score calibration process. In [32, 166], QMFs are formulated as additional components in the linear calibration strategy. In QMF, an additional quality term Q depends on reference and probe sample quality measures $\lambda_{reference}, \lambda_{probe}$:

$$S'_Q = w_0 + w_1 S + Q(\lambda_{reference}, \lambda_{probe}). \quad (6.5)$$

The QMF calibration is interpretable as a linear system fusion of the biometric comparison with a subsystem quantizing the quality of reference and probe. Likewise, one might reformulate the equation to calibrate LLR based thresholds regarding an updated prior belief in quality. Thereby, the system output S remains unaltered, whereas the threshold is altered, which could be compared to S'_Q , i.e., the threshold that equals S'_Q for a least-favorable decision¹²⁰:

$$S = w_1^{-1} (S'_Q - w_0 - Q(\lambda_{reference}, \lambda_{probe})). \quad (6.6)$$

Previous works [166, 167] introduce the following duration- and SNR-dependent QMFs, among others:

$$Q_{duration} := w_2 \log(d_{probe}), \quad (6.7)$$

$$Q_{SNR} := w_2 \text{SNR}_{probe}, \quad (6.8)$$

$$Q_{duration \& SNR} := w_2 \log(d_{probe}) + w_3 \text{SNR}_{probe}, \quad (6.9)$$

where d_{probe} , SNR_{probe} denote the duration and SNR of the probe sample, respectively. Reference samples are assumed to stem from the clean/full (noise free long utterance) condition.

QMFs sustain a parsimonious degree-of-freedom regarding the number of calibration parameters that needs to be trained. The assumption of this work is that the more calibration parameters are needed to train, the less robustness is provided by these calibration schemes in unknown conditions.

¹²⁰ Similar thought experiments hold for the presented cohort based score normalization. At the transition between comparison and decision subsystems, the last step in comparison can be reformulated as the first step of decision making.

6.2.2 Unified Audio Characterizations (UACs)

In [38], the proposed UAC calibration scheme utilizes a symmetric, bilinear combination matrix \mathbf{W} for conducting a quality similarity score between reference and probe UAC-/q-vectors $\mathbf{q}_{\text{reference}}, \mathbf{q}_{\text{probe}}$:

$$Q_{\text{UAC}} := w_2 \mathbf{q}_{\text{reference}}^T \mathbf{W} \mathbf{q}_{\text{probe}}. \quad (6.10)$$

The estimation of \mathbf{W} induces tremendously high degrees-of-freedom to the calibration stage, i.e., $\|\mathbf{q}\|^2$ parameters need to be estimated additionally. When considering all of the 55 examined conditions of this study, this would correspond to 3 025 \mathbf{W} parameters (additional to the one w_2 parameter). Low-rank estimates can be obtained by probabilistic principal component analysis.

In an extension of UAC [242], trial-based calibration is proposed for examining a set of 14 distinct conditions comprising cross-language, cross-channel, noisy, and reverberant effects. As such, [242] aims at a wide range of condition types. In this work, distinct condition types are examined. The trial-based calibration of [242] causes an immense high degree-of-freedom on large-scale operations as the most-alike samples for calibration training are selected for each comparison (approximately 500).

Following the QMF intention of parsimonious robustness against unseen conditions during calibration training, Q_{UAC} turns unfavorable when condition amounts increase. Thus, UAC based score calibration schemes using bilinear combination matrices are not pursued in this research.

6.2.3 Contribution: Function of Quality Estimates (FQEs)

Motivated by UACs and by QMFs, score calibration schemes are proposed here for q-vectors, namely **function of quality estimates (FQEs)**. Based on the q-vector design, the cosine distance similarity between reference and probe q-vectors $\mathbf{q}_{\text{reference}}, \mathbf{q}_{\text{probe}}$ is suitable as an FQE $Q_{\text{q-vector}}$, as alike conditions are linkable by cosine similarity (cf. Fig. 6.2b):

$$Q_{\text{q-vector}} := w_2 \cos(\mathbf{q}_{\text{reference}}, \mathbf{q}_{\text{probe}}). \quad (6.11)$$

Compared to the calibration model proposed for UACs [38], the proposed FQE requires far fewer calibration parameter estimations, i.e., one parameter w_2 . Calibration methods are sought to parsimoniously preserve robustness against unseen data. Since measuring certain quality metrics such as SNR is difficult in low-quality conditions, alternative calibration schemes relying on FQEs appear promising.

Baseline results of an uncalibrated system are shown in Fig. 6.6: performance in terms of C_{llr}^{\min} degrades significantly in lower SNR levels and on shorter observations. All conditions yield respectively

high C_{llr}^{mc} . The gap between C_{llr} and C_{llr}^{min} for SNR levels ≥ 15 dB is very small. Since similar trends on CROWD and AC noise are found, comparable to section 6.1.3, experimental results are only reported w.r.t. CROWD noise. The plot depicts log-log axes to emphasize two aspects: for low quality conditions, degradation impacts are more severe than for high quality conditions; and (more relevantly) for each condition, the calibration loss is not negligible considering the discrimination performance of each condition, independently of another. (For these purposes, an easier visual comparability of condition-wise performance is sacrificed.)

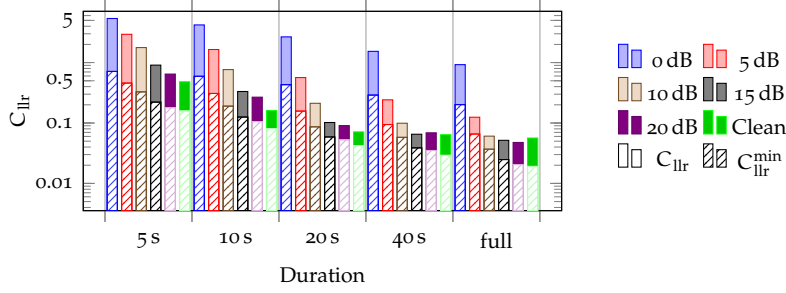


Figure 6.6: Baseline performance of uncalibrated system on CROWD conditions.

Fig. 6.7 shows the effects of conventional, QMF and FQE calibrations regarding the score distribution of genuine and impostor scores. Score distribution after QMF and FQE calibration resemble each other well. The conventional calibration appears to handle the full/clean condition very well, while suffering in other conditions. Opposed to conventional calibration, QMF and FQE calibration methods deal well with noisy and short probe samples. Notably, this behavior could as well arise due to the amount of training data from each condition present in the training process of calibration parameters. The training of conventional calibration is performed using only scores from the full/clean condition, whereas for training of QMF and FQE calibrations, scores from 55 conditions are utilized.

In the following study, comparisons are drawn between FQEs and QMFs that are employing the oracle SNR levels and the sample duration. In order to provide a compact overview, pooled conditions are examined as well, i.e., the performance of scores pooled from all examined conditions is evaluated at once instead of condition-wise. On the robustness of QMF and FQE calibration schemes against unseen conditions, evaluations are grouped into five analyses on segregated conditions known during calibration training: i) *duration impact on clean speech*, only conditions of 5 s, ..., full without noise are considered during calibration training and evaluation; ii) *SNR impact on full duration*, only noisy conditions with full duration (including the full/clean condition) are considered during calibration training and evaluation; iii) *combined duration and SNR effects*, all 55 conditions are

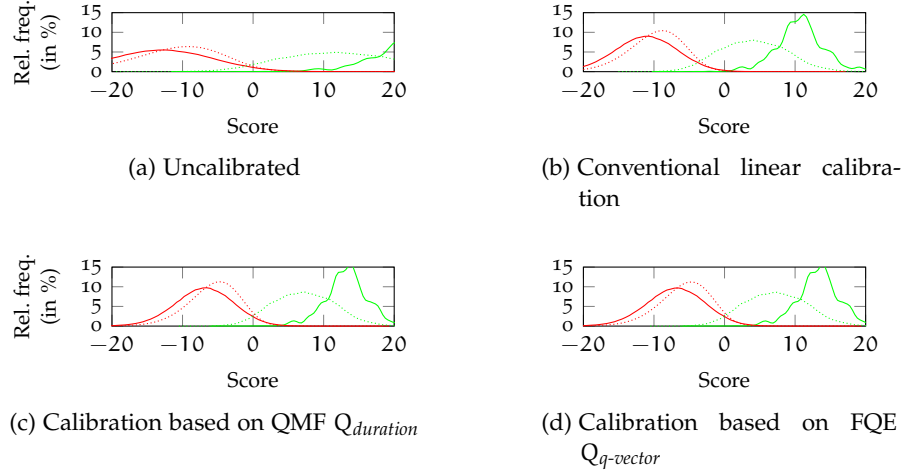
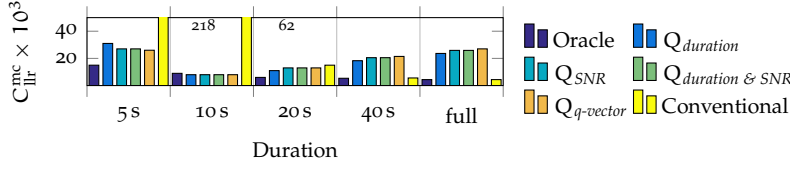


Figure 6.7: Comparison of class \mathcal{A} (green) and class \mathcal{B} (red) score distributions before and after calibration: full lines represent noise free long probe samples, dashed lines indicate scores from all conditions represented in Tab. 3.4 (excluding full/clean).

considered during calibration training and evaluation; iv) *pooled duration and SNR levels*, whereas i)-iii) examine each condition one-by-one and conditions are pooled (under equally prior conditions, i.e., at maximum uncertainty); and v) *calibration robustness to unseen conditions*, conditions that have durations and/or SNR levels of low quality (duration ≤ 10 s, SNR ≤ 5 dB) are excluded from training (training with 12 out of 55 conditions).

6.2.4 Analysis: Duration Impact on Clean Speech

This analysis focuses on noise free (clean) probe samples truncated into shorter durations. The calibration parameters for QMF and FQE are trained with recognition scores from noise free probe samples. Fig. 6.8 compares the three introduced QMFs to the proposed FQE (q-vectors), where *oracle* indicates ideal linear score calibration for a single condition (condition-wise calibration training, assuming ideal binning), and *conventional* indicates mismatched linear score calibration (solely trained on idealistic data of the full/clean condition). QMFs and FQE approaches perform better than conventional calibration on durations ≤ 40 s in terms of C_{llr}^{mc} . In terms of C_{llr}^{min} , no significant differences are observed between all examined calibration schemes. When the duration of probe samples is ≥ 40 s, these probe samples can be considered “full”, and the conventional calibration performs best.

Figure 6.8: C_{llr}^{mc} comparison of duration conditions (clean).

6.2.5 Analysis: SNR Impact on Full Duration

In this experiment, only noisy probe samples are drawn from noise free long duration (full/clean) samples. When training calibration, parameters for QMF and FQE are trained on probe samples that are not truncated. When training QMF calibration, applied SNR-levels are used instead of measured SNR (eradicating another source of biased performance results). This selection biases the performance of QMF calibration in different directions depending on SNR region. In the low-SNR region ($SNR \leq 15$ dB), estimating an SNR level is problematic due to almost equal levels of noise and speech present. Hence, the applied SNR level is much more accurate for the scope of this study than measured SNR levels. On the contrary, since voice data originally distributed for the [speaker recognition evaluations \(SREs\)](#) by the [US National Institute of Standards and Technology \(NIST\)](#) is seldom noise free¹²¹, the applied SNR level is less accurate compared to a measured SNR level for $SNR \geq 10$ dB.

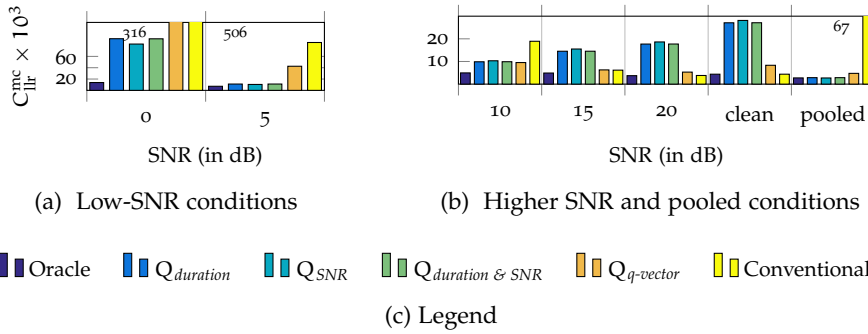
Figure 6.9: C_{llr}^{mc} comparison of different SNR conditions (full duration).

Fig. 6.9 depicts the calibration loss, C_{llr}^{mc} , across variable SNR but in full duration conditions. In the low-SNR region, QMF outperforms FQE based calibration in terms of miscalibration cost. This can be explained by the inclusion of the applied SNR-level as parametric values of QMFs. The quality estimates aid FQE calibration in handling low-SNR conditions compared to the conventional calibration. However,

¹²¹ The concept of *noise free* signals can be misleading, as noise is a mere summary term of undefinable and undesired signal characteristics—every real world signal conveys *noise* to some extent.

quality estimates are not fully accurate, since the applied SNR-level could be not as accurate itself, which, in turn, also undermines FQE calibration training. Dealing with $\text{SNR} \geq 10$ dB, FQE presents superior performance in terms of $C_{\text{llr}}^{\text{mc}}$ compared to QMFs. This is in line with the previous argument on SNR estimation in high-SNR, implying that quality estimates could be more accurate than the applied SNR level. In terms of $C_{\text{llr}}^{\text{min}}$, the proposed FQE reveals performance degrades, with relative losses to other calibration schemes of 23% to 36%, depending on the condition, while other calibration schemes yield almost similar $C_{\text{llr}}^{\text{min}}$, cf. Fig. 6.10.

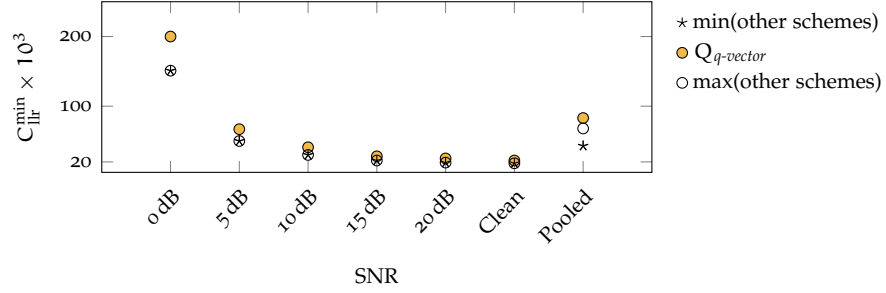


Figure 6.10: $C_{\text{llr}}^{\text{min}}$ comparison of $Q_{q\text{-vector}}$ on SNR conditions to the min./max. of the other schemes (oracle, Q_{duration} , Q_{SNR} , $Q_{\text{duration} \& \text{SNR}}$, conventional).

6.2.6 Analysis: Combined Duration and SNR Effects

Figs. 6.11, 6.12 and 6.13 depict $C_{\text{llr}}^{\text{mc}}$ losses for different quality ranges for the purpose of visualizing alike patterns in performance changes when moving across conditions. (Again, the visual comparability of single performance values is sacrificed for the purpose of visualizing dynamics in quality impacts; some merely noticeable differences between different calibration schemes are just too small to emphasize.) Thereby, for the sake of tractable analyses, the term *low quality* is referred to if either the duration ≤ 10 s or the SNR ≤ 5 dB. Other conditions are referred to as *good quality*. In general, no significant differences are observed between QMFs and FQE in terms of both $C_{\text{llr}}^{\text{min}}$ and $C_{\text{llr}}^{\text{mc}}$.

In Fig. 6.11, the miscalibration costs in conditions of short duration *and* SNR quality are visualized. The changes in miscalibration cost between 10 s and 5 s (at fixed SNR level) conditions are comparatively low compared to the changes between 5 dB and 0 dB (at fixed duration). QMFs and the proposed FQE perform similarly per condition and resemble a far better approximation of the oracle calibration scheme than the conventional calibration approach.

For the sake of visualizing QMF and FQE behavior in Figs. 6.12 and 6.13, values of $C_{\text{llr}}^{\text{mc}} > 0.1$ are cropped. Fig. 6.12 depicts the miscalibration cost in conditions of short duration *or* low-SNR quality. In

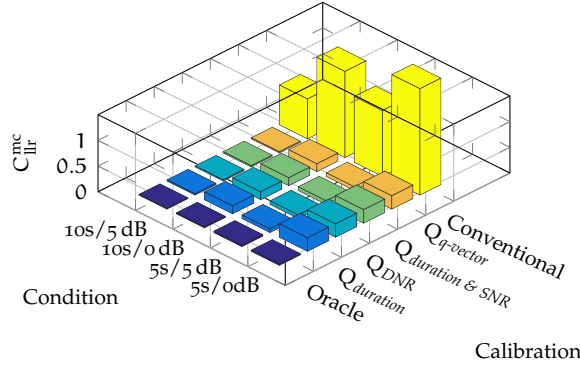


Figure 6.11: C_{llr}^{mc} comparison of combined low quality conditions.

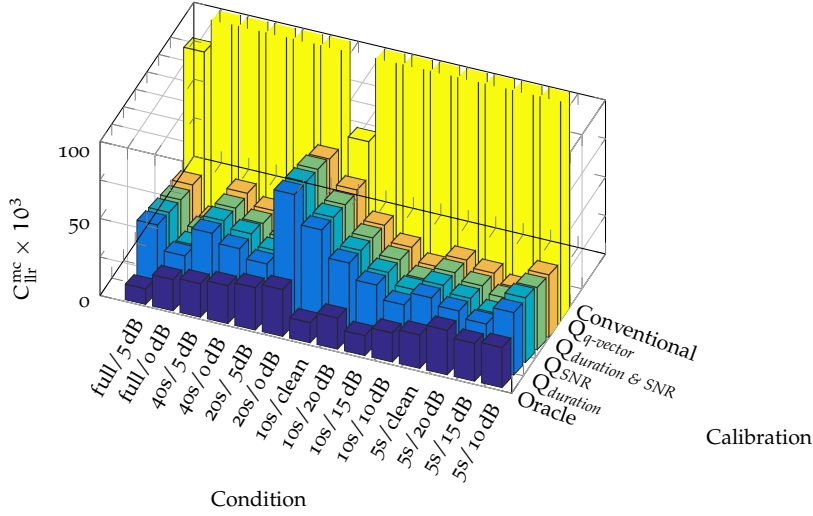


Figure 6.12: C_{llr}^{mc} comparison for having either short probe samples (≤ 10 s) or low-SNR (≤ 5 dB) present in the probe sample.

terms of approximating the C_{llr}^{mc} of the matched calibration scheme, QMFs and the proposed FQE outperform the conventional calibration scheme. For the 10 s/clean condition, quality based and the conventional calibration approaches yield somewhat similar results compared to the matched calibration scheme. The C_{llr}^{mc} of the matched calibration, however, increases with subsiding durations compared to (the lower) SNR level. By comparison, quality based calibration schemes are more severely affected by quality degradation in SNR than in duration. This observation is counter intuitive at first sight, since duration serves as a proxy measure to sample completeness, and shorter duration should indicate less information. On shorter durations, the discrimination performance (already) declines, see chapter 5; the miscalibration cost adds to recognition performance as discrimination and calibration. Calibration in conditions of shorter durations (of lacking information) is less (but still) difficult compared to calibration in conditions of lower SNR levels (of distorted information).

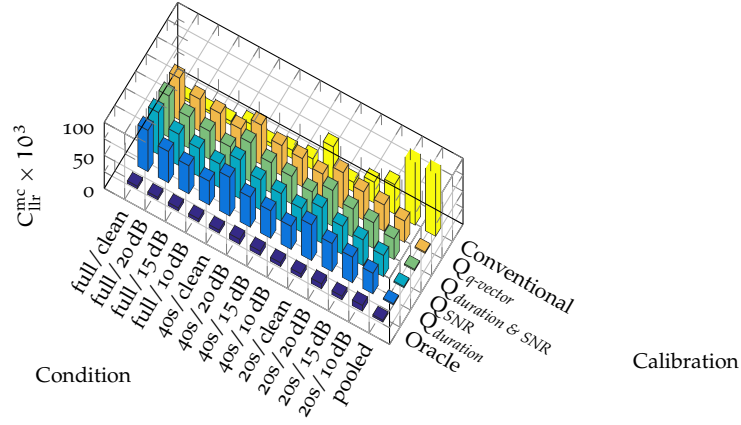


Figure 6.13: C_{llr}^{mc} comparison of good quality probe conditions (duration ≥ 20 s and SNR ≥ 10 dB) to pooled condition.

For good quality conditions, cf. Fig. 6.13, the miscalibration losses of QMFs and the proposed FQE increase. As the ideal condition of the conventional calibration scheme is reached, it approximates the oracle calibration much better than the examined quality based calibration schemes (in the vast majority of good quality conditions). Looking at the condition pooled C_{llr}^{mc} , however, the conventional approach is clearly outperformed by the QMFs and the proposed FQE, which approximate the calibration loss of the oracle calibration scheme well. From the above analysis, one might conclude that quality estimates can be successfully applied in the calibration stage providing similar levels of performance as the use of QMF calibration with oracle (applied) SNR levels and duration.

6.2.7 Analysis: Pooled Duration and SNR Levels

In Tab. 6.1, C_{llr}^{min} and C_{llr}^{mc} , performances of conventional, QMFs, and the proposed FQE calibration approaches are compared for pooled conditions w.r.t. variable duration on clean samples, variable SNR on full duration, and combined duration and SNR effects. On pooling variable duration conditions (pool D), QMFs and the proposed FQE yield very similar performances in terms of C_{llr}^{min} and C_{llr}^{mc} . By calibrating a pool of scores originated from different SNR levels in probe samples (pool S), however, the proposed FQE calibration approach fails to produce a better C_{llr}^{min} and C_{llr}^{mc} performance than the examined QMFs. The C_{llr}^{min} provided by the FQE $Q_{q-vector}$ is even worse than the C_{llr}^{min} of the conventional calibration. Such behavior could result from poor q-vector modeling over different SNR levels or could be partly attributed to the suitability limits of the cosine function in the FQE calibration scheme. When the whole range of duration and SNR variation is known during calibration training and evaluation (pool D+S), the calibration performance provided by $Q_{q-vector}$ is on

a par with the examined QMFs. It is deemed that supplying scores from combined duration and SNR variability to the training stage of calibration results in robust calibration parameter estimation.

Table 6.1: QMF and FQE comparison of pooled conditions: variable duration (D), variable SNR (S) and combined (D+S) conditions (in $C_{llr}^{min} \times 10^3$, $C_{llr}^{mc} \times 10^3$). As $C_{llr} = C_{llr}^{min} + C_{llr}^{mc}$, the comparison of C_{llr} values is trivial for these results.

Pool	Metric	Matched	$Q_{duration}$	Q_{SNR}	$Q_{duration \& SNR}$	$Q_{q-vector}$	Conventional
D	C_{llr}^{min}	59	70	68	68	67	70
	C_{llr}^{mc}	4	4	4	4	4	47
S	C_{llr}^{min}	43	67	68	67	83	68
	C_{llr}^{mc}	3	3	3	3	5	67
D+S	C_{llr}^{min}	126	160	160	160	159	160
	C_{llr}^{mc}	3	2	2	2	2	266

6.2.8 Analysis: Calibration Robustness to Unseen Conditions

From those previous observations on *low* and *good quality*, the robustness of unseen *low quality* deserves attention. Experiments towards robustness are conducted regarding pooled conditions but without knowing low quality conditions during calibration training, i.e., every condition afflicted by 5 s, 10 s or by 0 dB, 5 dB is excluded from calibration training. Thus, calibration functions are assumed to be solely trained on *good quality*. In evaluating calibration methods, the whole range of duration and SNR conditions (including *low quality*) is tested.

Table 6.2: QMF and FQE robustness comparison of pooled conditions with limited data during calibration training: variable duration (D), variable SNR (S), and combined (D+S) conditions (in $C_{llr}^{min} \times 10^3$, $C_{llr}^{mc} \times 10^3$), with differences to Tab. 6.1 in brackets.

Pool	Metric	$Q_{duration}$	Q_{SNR}	$Q_{duration \& SNR}$	$Q_{q-vector}$
D	C_{llr}^{min}	69 (-1)	69 (+1)	69 (+1)	67 (0)
	C_{llr}^{mc}	23 (+19)	23 (+19)	22 (+18)	19 (+15)
S	C_{llr}^{min}	68 (+1)	68 (0)	67 (0)	73 (-10)
	C_{llr}^{mc}	35 (+32)	36 (+33)	37 (+34)	38 (+33)
D+S	C_{llr}^{min}	160 (0)	160 (0)	160 (0)	159 (0)
	C_{llr}^{mc}	15 (+13)	17 (+15)	17 (+15)	14 (+12)

Tab. 6.2 provides the calibration performance for each experimental setup on pooled scenarios (evaluations are depicted separately in accordance with the three above pools D, S, D+S). In general, QMFs and

the proposed FQE are still outperforming the conventional scheme, although C_{llr}^{mc} costs have increased significantly: on pool D, the $Q_{q-vector}$ scheme yields the lowest C_{llr}^{mc} as well as the lowest C_{llr} . On pool S, C_{llr}^{mc} costs of the QMFs are slightly lower than on the proposed FQE. C_{llr}^{min} gains are observed for $Q_{q-vector}$. On combined conditions, pool D+S, C_{llr}^{mc} costs of the FQE are less affected than on the QMFs.

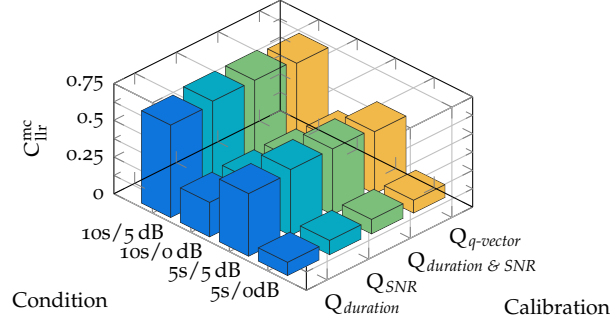


Figure 6.14: C_{llr}^{mc} comparison of combined low-quality, which is excluded from calibration training.

In order to gain more insight into the robustness of quality based calibration, more detailed analyses are performed that look into the individual duration/SNR conditions. Figs. 6.14, 6.15 and 6.16 depict the C_{llr}^{mc} costs of a calibration training without low quality conditions. Visual comparisons to oracle and conventional calibration schemes are spared (redundant information). Comparisons to results of the previous section are spared as well since the impact pattern of quality conditions is emphasized here for quality based calibration schemes. Changes in miscalibration costs are (nothing but) expected.

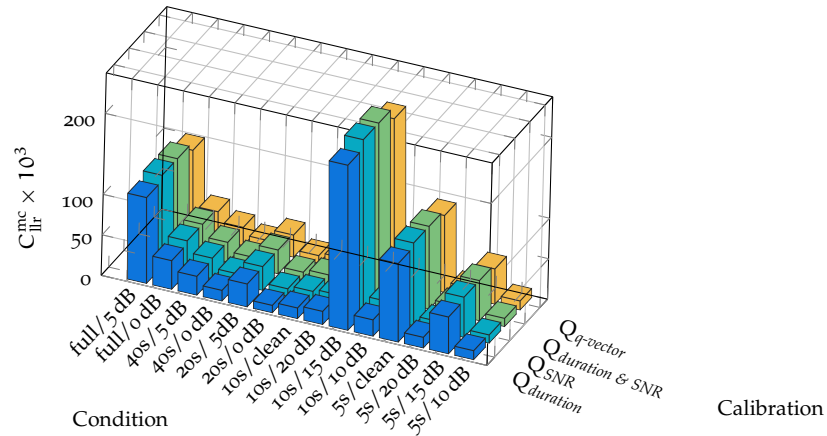


Figure 6.15: C_{llr}^{mc} comparison for having either short probe samples (≤ 10 s) or high level of noise ($SNR \leq 5$ dB) present in the probe sample with excluded low quality conditions from calibration training.

Miscalibration cost, as was expected, increases on low quality and decreases on higher quality conditions (as of the held back calibration training data). Nonetheless, QMFs and the FQE still outperform the conventional scheme on all low quality conditions. In contrast to knowing low quality conditions in training, the C_{llr}^{mc} pattern for low quality conditions reverses (cf. Fig. 6.14): 5 dB conditions suffer from much higher calibration losses than 0 dB conditions on short durations. Observing immense C_{llr}^{mc} increases for the four conditions of low quality in duration and SNR, the need of knowing low quality conditions becomes eminent. In conditions of mixed low and good quality (cf. Fig. 6.15), C_{llr}^{mc} losses of 5 dB conditions with full and 20 s durations are much higher than for their 0 dB variants. The same applies to the 15 dB conditions with 5 s and 10 s durations compared to the other short duration conditions, even if the 5 s/clean condition also yields comparatively high C_{llr}^{mc} costs across all quality based calibration schemes. The calibration of unknown low-SNR conditions in the midst of full durations results in better calibration performance than the calibration of unknown short duration conditions at mid noisy to clean speech. On testing with good quality data (cf. Fig. 6.16), the performance of both QMF and FQE calibration schemes occasionally falls behind the conventional approach. In contrast, for calibration training with low quality conditions, the C_{llr}^{mc} costs reduce with lower quality conditions. Here, worse calibration performance is observed.

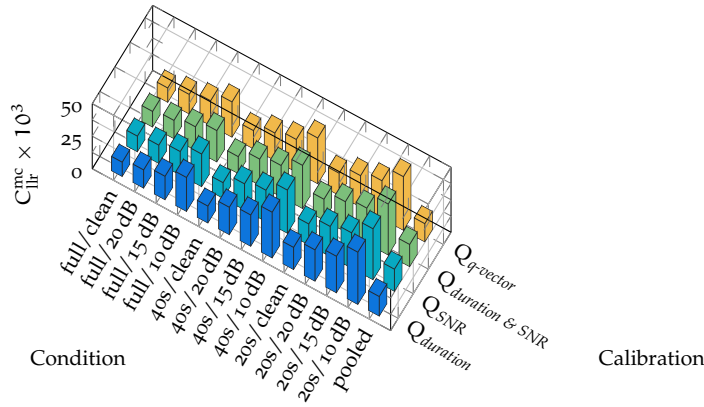


Figure 6.16: C_{llr}^{mc} comparison of good quality probe conditions (duration ≥ 20 s and SNR ≥ 10 dB) to pooled condition with excluded low quality conditions from calibration training.

The present analysis provides insights into the behavior of calibration schemes in combined conditions of high signal degradation and short segment duration regarding accurate approximation of idealized linear calibration. QMFs and the proposed FQE reduce C_{llr}^{mc} costs down to 5% to 6% of the conventional calibration scheme if all conditions are known; and down to 10% to 12% in the presence of unseen low quality conditions. Quality based calibration schemes are essen-

tial when facing unconstrained quality conditions since they are more robust towards unknown conditions than the conventional condition-mismatched calibration approach. A proper function is required to account for single-valued conventional quality measures or vector-based quality estimates in the calibration process. The performance of q-vector based calibration using cosine function is investigated, leaving analyses towards more efficient techniques for other FQEs to further research. The current study indicates that quality based calibration provides more reliable calibrated scores, especially when confronted with samples of low quality conditions.

6.3 DEEP SCORE NORMALIZATION WITH QUALITY ESTIMATES

The previous sections dealt with the impact of q-vectors to score normalization. At first, a quality based cohort pre-selection was examined, a linear score calibration based on quality estimates was assessed afterwards. The proposed score normalization scheme resulted in better discrimination performance, and the proposed calibration scheme resulted in lower calibration losses.

In this section¹²², emphasis has been placed on a deep neural network (DNN) scheme to examine the robustness of quality informed score normalization regarding SNR levels and noise types. The DNN is not proposed here as a solution of some sort but for assessing the limitations and possible gains of preceding studies. Eventually, conclusions are drawn by the following DNN study to further improve the existing conventional and proposed methods. Depending on the DNN design, its output¹²³ can be assumed as *well-calibrated* (regarding the empirical prior $\tilde{\pi}$ of the application-dependent training data set, see sections 2.1 and 2.4.4). By contrast, this study solely concerns the discrimination performance since the mutual optimization of discrimination and calibration performance (without following the LLR paradigm by design) might lead to distorted performance (an application-dependent *coin tossing* system which has no calibration loss would still not be discriminant). In other words, solely targeting a low calibration loss regarding the production of LLR scores is not feasible when exploiting discrimination power by utilizing neural networks. This study therefore solely concerns discrimination performance to the benefit of sustaining lower computational efforts but at the risk of higher calibration losses.

In this study, the input to a deep neural network comprises reference and probe i-vectors, corresponding q-vectors, and depending

¹²² Parts of this section are based on a collaborative work with Søren Trads Steen and Christoph Busch [72] which emerged from the collaboration with Søren on his master's thesis [247].

¹²³ By the conventional use of *softmax* end layers, the output of neural networks is meant to correspond to posterior probabilities, the generation of prior independent outputs being LLRs lies outside the scope of this dissertation.

PLDA scores. The DNN is trained using the I4U development set (here, the cohort data).¹²⁴ Noise levels and types are thereby removed from the network training in order to gain insights into depending score normalization functions in a fair evaluation setup to be compared to preceding results. The expectation is that deep score normalization schemes can yield discrimination gains. However, when removing cohort data of low quality or of noise types, performance losses are expected. Robustness is addressed in terms of whether or not a soft relative bound can be sustained. In other words, although the input data to the deep neural network has all information about the conditions available, a re-disposition of the network training task emulates a quality domain shift scenario. The following analyses will thus investigate on how well the (oracle) discrimination performance can be sustained, despite an ill-disposition of the network training task, given rich input data to the network.¹²⁵ In this section, two hypotheses are investigated: i) as a single neuron (network) mathematically equals linear score calibration, and the proposed FQE calibration sustains discrimination performance, a deeper and more elaborate network architecture can achieve gains in C_{llr}^{\min} across conditions; and ii) on changing the disposition of the (oracle) score normalization training, i.e., by excluding cohort data comprising conditions representing low quality from training (short duration and low-SNR of both noise types versus all SNR levels of the biometric noise type), C_{llr}^{\min} discrimination performance will maintain within a $\pm 20\%$ band.¹²⁶

6.3.1 Conventional Deep Neural Networks

A conventional feed forward neural network is employed with softmax training scheme. Feed forward neural networks consist of layers of units [58]: an input layer and an output layer are linked over a number of hidden layers by numerous connections. The connections between units of each layer are weighted. Weights are initialized with small numbers in terms of a $\sqrt{2/n_l}$ standard deviation [248], with

¹²⁴ To account for European data privacy regulations [55] has been partly motivated by the collaborative work of [72]. The data privacy of cohort data subjects is preserved as aggregated information and distributed in the form of network parameters rather than distributed as voice samples of each cohort speaker. As such, data deletion procedures would be easier to enable if cohort subjects revoked their data privacy consent in the future (after network training).

¹²⁵ As *quality* is a vastly broad area in speech processing and speech communication, one may utilize quality estimates of other areas instead.

¹²⁶ The 20% relative band is arbitrarily set. The value represents a (soft) measure for *robustness due to the lack of data* (considering that real world data will differ from training data) for recognition systems [24, 29] *to maintain the tendency of its performance when reducing the quality conditions of the data under examination*. As any relative band would be arbitrary, 20% are chosen, representing an upper bound for the purpose to claim *robustness*. In other words, the aim of requiring a 20% relative band is to rather reject robustness claims than to confirm them.

n_l representing the number of incoming connections to the unit. In each unit, the *response* is constructed as a linear combination of the outputs from the units of previous layers. A non-linear activation function is evaluated on the response to achieve the output (*activation*) of the units by, e.g., the rectified linear unit (ReLU) activation function [249] and the sigmoid function for bounded activations [58]. Networks are trained to optimize performance regarding the binary cross-entropy cost function using gradient descent, where the Adam algorithm [250] and *backpropagation* [58] are employed. A single-unit output layer is utilized, representing a system's normalized score. In order to avoid over-fitting of the training data, different regularization schemes can be employed, such as *weight decay* [58], *dropout* [251], and *batch normalization* [252].

6.3.2 Contribution: Deep Quality Informed Score Normalization

Accounting for quality information as well as cohort-related data, a feed forward neural network is constructed based on the comparison score, reference and probe i-vectors \mathbf{i}_{ref} , \mathbf{i}_{prb} as well as corresponding q-vectors \mathbf{q}_{ref} , \mathbf{q}_{prb} as the concatenated input vector, cf. Fig. 6.17. A normalized score between 0 and 1 (representing classes \mathcal{B} and \mathcal{A}) is obtained via a single unit output layer with a sigmoid activation function. By training the network on the cohort dataset, the network model is assumed to comprise cohort and quality information, while achieving anonymity (not only pseudonymization) for the cohort speakers. For the purpose of accounting for linear normalization approaches, e.g., linear quality calibration [38, 71, 79], the first hidden layer of the proposed network employs a linear activation function $f(x) = a + b x$. During training, input features are adaptively normalized w.r.t. the amount of genuine and impostor comparisons. Deeper hidden layers are connected by non-linear activation functions, e.g., using the ReLU activation function. In order to achieve an effective class balance of equal priors, genuine comparisons are weighted higher than the impostor comparisons during network training. The network configuration is referred to as (L, U) with a network of a linear layer with U units, followed by L non-linear layers of U units, cf. Fig. 6.17.

In order to examine network configurations¹²⁷, $L = 1, 2, 4$ layers are investigated. All layers comprise the same amount of hidden units, i.e., $U = 50, 100, 200$ units. Tab. 6.3 compares the different networks on the test set: configuration (1, 50) yields the largest condition-average C_{llr}^{\min} gain over a conventional i-vector/PLDA baseline system of 6.2%

¹²⁷ Convergence is reached after 3 epochs on a randomly selected 20% held-out validation subset on which the best performing model is chosen. On a fixed number of layers and units, a learning rate of 10^{-5} is found to reduce over-fitting well on the hold-out validation set.

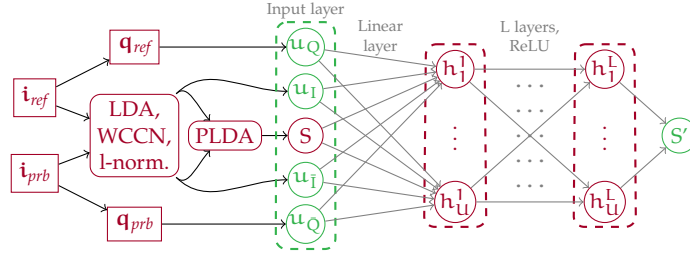


Figure 6.17: Deep score normalization architecture with quality estimates.

The baseline system comprises a standard i-vector acoustic feature extraction. Then, i-vectors are processed by linear discriminant analysis (LDA), within-class covariance normalization (WCCN), and length normalization (l-norm.), resulting in biometric i-vectors (the input units u_I, u_U). As i-vectors are compared by PLDA, resulting in a score S , this score is considered as input as well. q-vectors q_{ref}, q_{prb} are extracted from acoustic reference and probe i-vectors i_{ref}, i_{prb} , serving as q-vector input units u_Q, u_Q to the DNN. The DNN has L layers and h_1^1, \dots, h_U^L hidden units (per layer), resulting in a normalized score S' .

with the lowest standard variation, i.e., with rather stable improvements among all conditions. Configuration (2, 100) yields the second largest gains regarding average and deviation in terms of C_{llr}^{\min} , but also regarding pooled-condition performance, where the (2, 50) network yields the largest gains. Accounting for potential over-fitting, dropout is examined on (1, 50) and (2, 100) networks with a 20% dropout rate: on average, C_{llr}^{\min} grows, which may occur due to an excessive dropout rate. By contrast, the cohort normalization in [70] yields up to 8.2% relative gains in C_{llr}^{\min} on single conditions, see section 6.1.3. Further investigations are carried out on the (1, 50) and the (2, 100) configurations, motivated by their gains in average performance across conditions and in the condition pooled performance.

Table 6.3: Benchmark of relative C_{llr}^{\min} changes (in %) to PLDA baseline on the test set regarding condition averaging (μ), standard deviation (σ), pooling (p), and dropout training (DO).

(L, U)	(1, 50)	(1, 100)	(1, 200)	(2, 50)	(2, 100)	(2, 200)	(4, 50)	(4, 100)	(4, 200)	(1, 50)	(2, 100)
DO											
μ	-6.2	1.4	-2.0	-2.1	-5.7	0.8	-2.9	-5.2	-0.9	1.3	4.9
σ	2.4	6.6	3.5	4.3	2.6	4.4	3.1	2.9	3.8	1.6	4.0
p	-4.6	0.9	-0.2	-6.6	-6.4	0.4	-0.2	-3.4	0.0	-2.5	7.1

6.3.3 Robustness Analysis: Contained SNR Levels and Noise Types

For the purpose of examining the robustness of the proposed normalization, training is conducted under two different test setups of (simulated) unseen data:

- Targeting low quality, all conditions afflicted with SNR levels ≤ 5 dB and with durations ≤ 10 s are excluded.
- Targeting the biometric noise type, all CROWD noise conditions are excluded. In contrast to AC noise (ambient noise), CROWD noise is biometric interfering noise. CROWD noise is derived by mixing the speech of other biometric subjects in an overlapping fashion.¹²⁸

Fig. 6.18 compares the effects to (1,50) and (2,100) configurations, with and without employing dropout, regarding whether or not the C_{llr}^{\min} performance is not exceeding a $\pm 20\%$ performance band w.r.t. the C_{llr}^{\min} performance of the (comprehensively trained) DNN in each condition. In the low quality analysis depicted in Figs. 6.18a, 6.18b, the (1,50) configuration outperforms the (2,100) in terms of robustness. Also, employing dropouts sustain coherent and stable performance, providing a rather robust network parameterization when the network is trained without low quality conditions. By focussing on robustness of noise type, both configurations perform stable and coherent results with slight benefits if dropout training is conducted, as illustrated in Figs. 6.18c, 6.18d. Expectedly, all good quality conditions benefit when low quality conditions are excluded from network training.

Examining an exemplary deeper architecture, gains compared to the baseline PLDA performance are observed. The use of the two layer configuration does not increase the performance of the one layer configuration much further. In the robustness analysis, i.e., by excluding low quality conditions and the more challenging noise type (CROWD noise), the proposed approach reveals to benefit on good quality conditions. The performance of the (1,50) configuration is preserved within a $\pm 20\%$ performance band¹²⁹ on unseen low

¹²⁸ An equivalent for overlapping biometric noise (as speech signals from TVs running in the background or the so-called cocktail party effect) are latent fingerprints at border control sensors that overlap with latent fingerprints from previous [biometric capture subjects](#) if the sensors are not cleaned. In that analogy, AC noise would rather represent the finger pressure on a sensor.

¹²⁹ In other words, the visualization of the 20% relative band in Figs. 6.18a and 6.18b directly leads to the conclusion that, without knowing any short duration or low-SNR conditions, the neural network is not robust. However, these conditions are known to the setup in Figs. 6.18c and 6.18d. All CROWD noise type conditions are removed, and all observations are observed within a 20% relative band. One can thus conclude that knowing the degree of quality degradation is more important than knowing various noise types, i.e., removing CROWD noise from the training does not decrease performance by more than 20%.

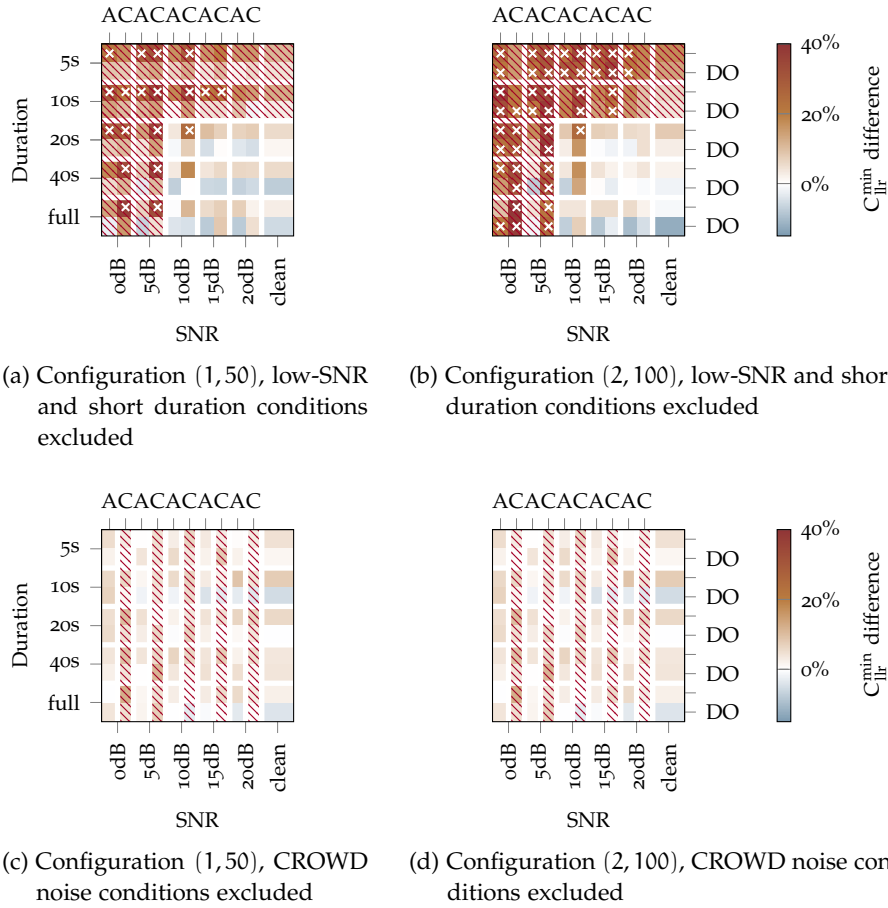


Figure 6.18: Relative C_{llr}^{\min} change on test set (in %). Performance by duration and SNR regarding AC (A) and CROWD (C) noise as well as whether dropout is conducted (DO). Red areas indicate the conditions excluded from training. Crosses denote relative C_{llr}^{\min} changes above $\pm 20\%$.

quality conditions. By contrast, when excluding overlapping speech (CROWD noise) conditions, both (1,50) and (2,100) configurations perform comparatively stable. Thus, the proposed approach rather benefits from training on a broad scale of SNR levels than on more noise types. This poses a challenging scenario due to overlapping biometric features of other subjects.

6.4 SUMMARY

In this chapter, score normalization and calibration methods were proposed to cope with performance losses in unconstrained environments. Thereby, the acoustic and biometric feature extraction as well as the biometric comparison remained fixed—a state-of-the-art i-vector/PLDA system that is trained in a condition pooled fashion.¹³⁰

¹³⁰ Oracle VAD is assumed, motivated by the observations in Chapter 5.

These biometric subsystems, however, are also considered *white boxes* as their signal processing is readable.

Based on acoustic features, **q-vectors** are proposed to characterize quality conditions of audios. Motivated by **UAC** [38], a q-vector represents the posterior probabilities of each condition for a given acoustic i-vector, i.e., from the signal processing perspective of the speaker recognition system. Quality conditions are sampled by log-linear performance observation, as performance changes are observed to decline with logarithmic quality degradation. The goal is to exploit the behavior of the signal processing subsystem to improve the scores communicated from the comparison to the decision subsystem (see section 2.2), such that discrimination and calibration performance are improvable by **biometric system owner**, **operator**, and **provider** (as well as by **vendors**).

Regarding the linkage and classification of conditions, q-vectors are beneficial: linkage is achieved by the cosine of two q-vectors and by the symmetric **Kullback-Leibler divergence**. Classification is achieved by finding the maximum posterior element (here, the maximum value in one q-vector). The classification of conditions, however, is not investigated in this dissertation. Condition classification is useful when it comes to splitting the signal processing into subsystems depending on the expected performance—by binning probe comparisons, scores are put into different sets.¹³¹

Based on q-vectors, a novel pre-selection to cohort based score normalization is proposed, targeting discrimination performance. Conventionally, cohort data comprises data of idealistic quality (from high quality conditions). When facing *unconstrained environments*, multiple other conditions need to be considered as well. Therefore, cohort data is also synthetically degraded (in the same manner). For pre-selecting the relevant cohort data, cohort data of alike q-vectors is found by the symmetric Kullback-Leibler divergence (proposed method), i.e., the cohort data that is similar to the biometric probe (which is of unconstrained quality).

Especially in low quality conditions, the proposed scheme increases the discrimination performance. The proposed normalization scheme

¹³¹ In score calibration by **pool adjacent violators algorithm (PAV)** (section 2.4.4), scores resembling alike LLRs are grouped. As this pinpoints the assessment of defined requirements (see verbal scales, chapter 4), the variety of LLR groups determines the amount of different (and meaningful) decision policies that could be employed by using a system. Here, a *meaningful manner* means that not all policies share the same threshold value and their differences may be interpreted within formally denoted decision requirements. For each of these conditions, score subsets are created—here, 55 subsets—and as score calibration is algorithm dependent, i.e., on each condition bin, LLR groups are calibrated for a 55th of the overall number of available scores. The variety of verbal bands a system is capable of supporting might be reduced drastically. Furthermore, by binning conditions, binning errors are introduced (when samples are wrongly assigned to conditions), which would require further treatment. In the context of this dissertation (for these reasons), if condition binning is applied, it should target the bare minimum of bins necessary.

re-biases and scales PLDA scores adaptive to the quality of a probe sample (from the perspective of the signal processing subsystem).

Targeting calibration performance, FQEs are introduced. In FQEs, q-vectors are linked in order to adaptively re-bias weighted PLDA scores. Exemplarily, the cosine similarity is proposed as an FQE. In comparison to QMFs—proposed in related work, based on proxy measures to quality (duration and SNR)—, FQEs estimate quality from the perspective of the signal processing subsystem. FQEs bridge the gap on the outline of the term *quality* between the perspective of task-independent speech quality (here: proxy measures to quality) and the perspective of a task-dependent utility predictor (biometric quality). This is done by encouraging the perspective of a task-dependent speech quality estimator (that serves to improve decision making, even if utility might be low). Generalizing over duration and noise conditions, the proposed FQE achieves a slightly better calibration performance than the compared QMFs. A detailed robustness study, however, motivates to employ conventional calibration schemes on good quality (neither QMFs nor FQEs), i.e., when the duration ≥ 40 s and the SNR level ≥ 15 dB. Otherwise, either QMFs and the proposed FQEs yield similar results in discrimination as well as in calibration performance. Originally, UACs (basis to q-vectors and thus to FQEs) were also proposed for score calibration: UACs are linked by a bilinear combination matrix. Employing bilinear combination matrices to q-vectors, however, a number of additional calibration parameters growing quadratic with the amount of conditions are implied. QMFs and FQEs sustain a minimalistic degree-of-freedom regarding the amount of calibration parameters necessary to be estimated. These calibration methods have a parsimonious degree of freedom (sustain low complexity in their amount of parameters) and are therefore assumed to provide higher robustness.

When reformulating the conventional and proposed score calibration schemes in terms of neural networks, they effectively resemble single neuron networks. Although the cohort score normalization study yields gains in C_{llr}^{\min} , no gains in discrimination performance are observed for score calibration methods. Thus, an exemplary deep neural network scheme is investigated regarding potentials and limits of q-vector based score normalization (and calibration) methods to discrimination performance. Calibration performance is not explicitly targeted; benefits in C_{llr}^{\min} result from employing a rather complex score normalization scheme. Moreover, a robustness analysis on the employed training data illustrates that one might consider to train score normalization on a variety of SNR levels rather than training score normalization on a variety of noise types (such analyses could spare future efforts).

To sum up, quality mismatches can be estimated on pre-comparison stages (from acoustic features) to aid Bayesian identity infer-

ence by enhancing conventional score normalization and calibration methods. Thereby, the proposed unconstrained cohort normalization scheme yields very promising results compared to conventional approaches. The proposed FQE calibration scheme achieves competitive results to the related work on QMFs. Meanwhile, the study on deep score normalization employing quality estimates revealed further discrimination gains and robustness towards parsimonious training schemes, detaining conditions of low quality and biometric-interfering noise types.

PRESENTATION ATTACK SECURITY, PRIVACY, AND DATA PROTECTION

This chapter addresses security as well as privacy and data protection, concentrating on the applicability of [log-likelihood ratios \(LLRs\)](#) and the [Bayesian decision framework \(BDF\)](#) in the real world. It particularly depicts research progress on i) [presentation attack detection \(PAD\)](#) security for short term replay attacks and on ii) privacy and data protection, targeting the research questions:

As security systems are subject to attacks, can text-independent audio replay attacks which are based on unit-selections of previously captured and rearranged speech sequences be detected?

Can data privacy and data protection be preserved while sustaining performance? On privacy and data protection:

- Within the framework of the 2016 (EU) *General Data Privacy Regulation* [55], considering biometric data as *sensitive*, is privacy preservable for [biometric capture subjects](#) while sustaining performance?
- Can the data of comparison subsystem vendors be protected, meaning parameters of generative comparison models, that are trained on vast data amounts, while preserving privacy of capture subjects and sustaining performance?

Modern text-to-speech algorithms, when employed for subversive usage (to generate [presentation attacks](#)), pose a vital threat to the security of speaker recognition systems. In order to distinguish between [attack presentations](#) and [bona fide presentations](#), the use of PAD subsystems is of utmost importance, see section 2.2. To this day, the vast majority of introduced spoofing countermeasures in speaker recognition has relied on speech production and perception based features. In this chapter, emphasis regarding PAD security is put on an audio replay attack, namely *unit-selection*. The classification of natural versus non-natural speech transitions is carved out based on (unfiltered) wavelet and Fourier frequency features.

Furthermore, data privacy is crucial when dealing with biometric data. Considering the 2016 European general data privacy regulation [55] and the second payment service directive [253], biometric information protection is essential for any commercial application. Biometric information is protected [44], when i) *unlinkability* is ensured

across [biometric system owners](#), [operators](#), and [providers](#), ii) *irreversibility* of encrypted information is sustained, and iii) *renewability* is guaranteed of, e.g., voice representations following the [intermediate-sized vector \(i-vector\)](#) paradigm. Biometric voice-based systems need to be prepared in compliance with the latest EU data privacy legislation. Therefore, LLR score properties need to be preserved in order to sustain a wide application range, seeking biometric information protection without performance loss. The protection of model parameters is also of interest for system [biometric system vendors](#), such that owners, operators and providers are fully unaware of the actual values involved in data processing (the model parameters of the comparison subsystem) as well as the data processed (the biometric information).

7.1 DETECTION OF UNIT SELECTION ATTACKS

In security scenarios, the performance of a biometric system is examined regarding subversive usage w.r.t. different system levels [254]. Attacks at the sensor level are referred to as presentation attacks [93]. Speaker recognition systems are particularly threatened due to the advanced development of speech synthesis techniques [49]. Voice presentation attacks are classified into six attack types [255]: synthesis [256], voice conversion [255], mock-up [257], replay [258], unit-selection [49], and mimicry [259]. Fig. 7.1 provides an overview on the different types of presentation attacks. In a speech synthesis attack, attackers create a synthetic voice of the targeted identity in order to synthesize speech samples which are accepted by the speaker recognition system [256]. In a voice conversion attack, an existing speech sample of the impostor is altered, such that it becomes more similar to the voice signal of the target subject [255]. In a mock-up attack, the impostor generates a synthetic signal in order to circumvent speaker recognition systems by causing high comparison scores that do not necessarily contain speech signals [257]. Replay attacks refer to the playback of a previous captured voice sample to the speaker recognition system [258]. For unit-selection attacks, speech samples of the attacked subject are captured, segmented into parts, called units, and replayed in different sequences to the speaker recognition system. Imitation or mimicry is the attempt of a [subversive user](#) to mimic an enrolled [subject](#) in order to get access to the system via the foreign account [259].

This section concerns¹³² a certain type of presentation attacks, namely unit-selection. Unit-selection [presentation attack instruments](#)

¹³² Parts of this section are based on a collaborative work with Ulrich Johannes Scherhag, Christian Rathgeb and Christoph Busch [73], which emerged from the collaboration with Ulrich over his master's thesis [260] (received the 2016 CAST IT-Security award with first place distinction among six finalists).

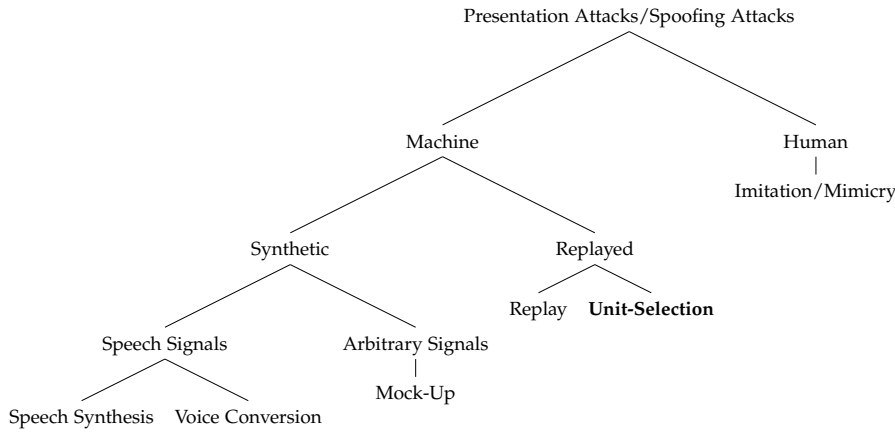


Figure 7.1: Structure of presentation attacks.

(PAIs) can be thought of as a collage of speech signals, introducing non-natural **presentation artefacts**, such as abrupt changes in frequency transitions. In this section, the following hypotheses are investigated:

- Collages of sound units comprise non-natural speech transitions, recognizable as edges by wavelet and Fourier based frequency features.
- Unit-selection attacks can be detected language-independently.

7.1.1 Voice PAD: the ASVspoof 2015 Challenge

In the ASVspoof 2015 spoofing challenge, the detection of text-independent attacks, in particular on voice synthesis, voice conversion, and unit-selection [49] is being focused on. Five out of ten attack algorithms are known during the PAD system development phase in order to investigate the detection robustness against (the other five) unknown attacks. State-of-the-art voice PAD systems [261–264] achieve **equal error rates (EERs)** of about 0% for each of nine out of ten **presentation attack instrument species (PAISs)**, i.e., for all PAISs of the ASVspoof 2015 challenge but the unit-selection as the challenge’s tenth PAIS (referred to as *S10*, the tenth spoofing attack). The unit-selection PAD performance resembles the *attack potential* of all of the challenge’s PAIS. The countermeasures utilize phase-based features, which detect non-natural phase shifts in generated artefact samples. During the synthesis process, only amplitude information is concerned in the vocoding stage, making phase-based features convenient for detecting such artefacts. In contrast to synthetic speech signals, unit-selection creates artefacts by reusing previous recorded samples [265]. The natural phase-shift of the sample is thus preserved and the applicability of countermeasures utilizing phase-based features is limited.

However, a successful countermeasure against unit-selection attacks, as proposed in [266], employs a feature-combination of cochlear filter cepstral coefficients (CFCC), instantaneous frequency (IF), and [mel-frequency cepstral coefficient \(MFCC\)](#). The CFCC, introduced in [267], is calculated by utilizing an auditory transform (AT), followed by a filter bank and discrete cosine transform (DCT). The AT itself is a function emulating the filter function of the cochlear [268]. In order to consider phase information, a CFCCIF is designed, combining CFCC with IF. Fused with MFCCs, this approach yields an EER of 1.2% on the ASVspoof data and an EER of 8.5% on unit-selection attacks, which is the best result achieved in the context of the ASVspoof 2015 challenge [266]. In [269], the same authors propose a unit-selection detection utilizing prosodic features, i.e., fundamental frequency (f_0) contour and strength of excitation, achieving an EER of 12.4% on the ASVspoof 2015 data. Eventually, constant Q cepstral coefficients (CQCCs) [144] are proposed.¹³³ CQCCs are proposed in music processing to analyze different frequencies with variable resolution. For unit-selections, CQCCs yielded a 0.5% EER.

Most common features that analyze frequencies, such as MFCCs and CFCCs, aim at emulating the perception of humans. However, the human hearing is rather specialized for speech recognition. For that reason, state-of-the-art presentation attack countermeasures are capable of yielding significantly better PAD performances compared to human observers [270]. On the contrary, CQCCs are successful as their processing is motivated from music analysis. Here, features are proposed based on artefact observations of unit-selection attacks.

7.1.2 Contribution: Countermeasure on Sound Unit Transitions

Examining the frequency-domain of unit-selection attacks, speech is interpreted as a concatenation of phonemes or likewise sound units. Concatenation points are referred to as transitions. Differences in the frequency domain of unit transitions in [bona fide presentations](#) and [attack presentations](#) are thereby exploited.

7.1.2.1 Frequency Analysis of Sound Unit Transitions

In bona fide presentations of (human speech), phonemes transition smoothly into another. The continuous transition of a bona fide speech signal is depicted in Fig. 7.2. Audio-signals, which are a compound (like a collage) of multiple voice fragments (phonemes or other units) and not smoothed afterwards, show more abrupt changes of the frequency in the signal, as illustrated in Fig. 7.3.

¹³³ The ASV spoof 2015 challenge leads to the proposal of different features. The CQCC paper on the ASVspoof 2015 challenge and the collaboration this work [73] is partly based on were in review/under submission simultaneously. CQCCs became the baseline for the 2017 and 2019 ASVspoof challenges.

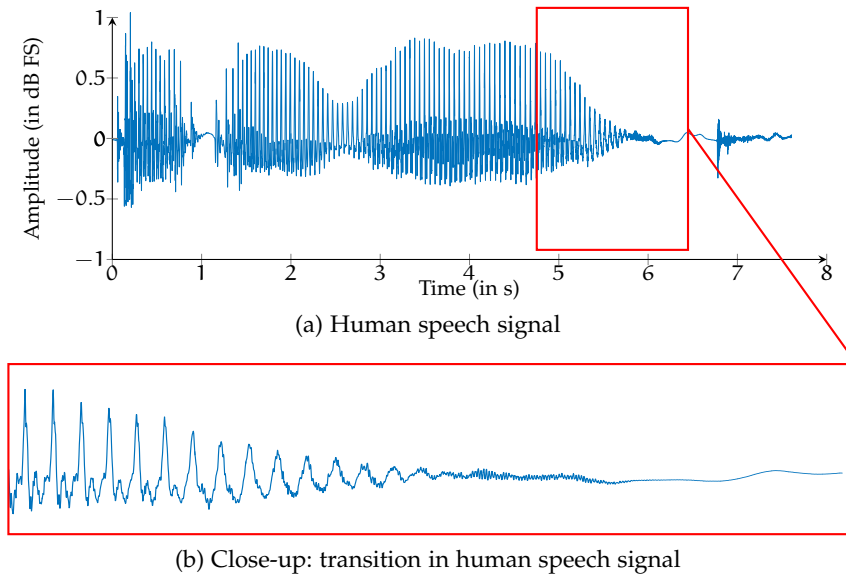


Figure 7.2: Example of a bona fide (human) speech signal and transitions.

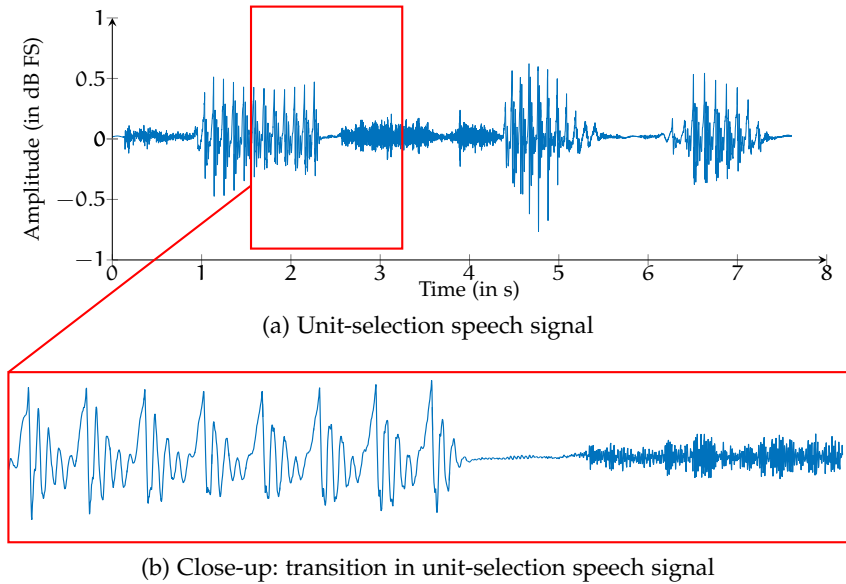


Figure 7.3: Example of a unit-selection speech signal and transitions.

In bona fide presentations, smooth transitions result in natural transitions in the frequency-domain as exemplarily depicted in Fig. 7.4, whereas the transformation of the non-natural concatenated signal causes abrupt changes in the whole frequency band.

Fig. 7.5 illustrates unit-selection [presentation artefacts](#) in the spectrum: higher frequencies comprise abrupt changes in the magnitude, which, compared to natural human speech, comprise more density and occur more often. Motivated by this analysis, Fourier and wavelet based features are proposed in order to distinguish between (bona fide) speech and unit-selection attacks.

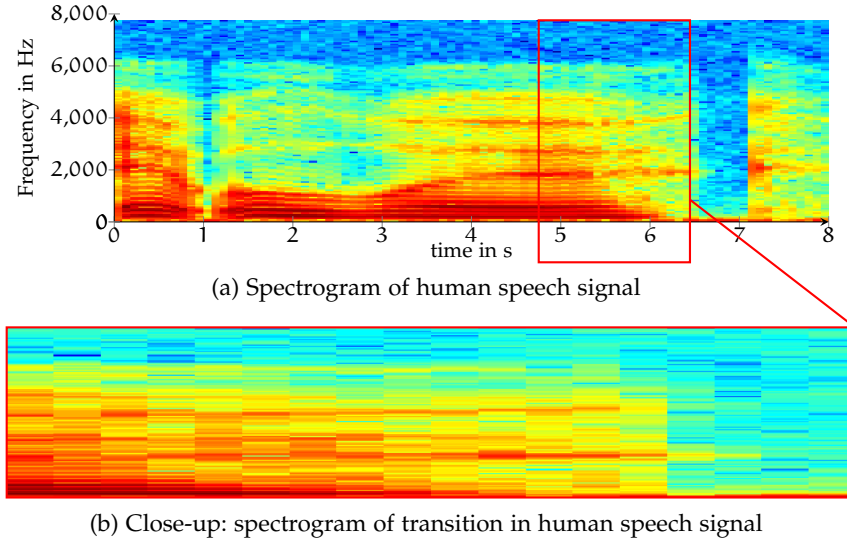


Figure 7.4: Spectrogram of a human speech signal and transitions.

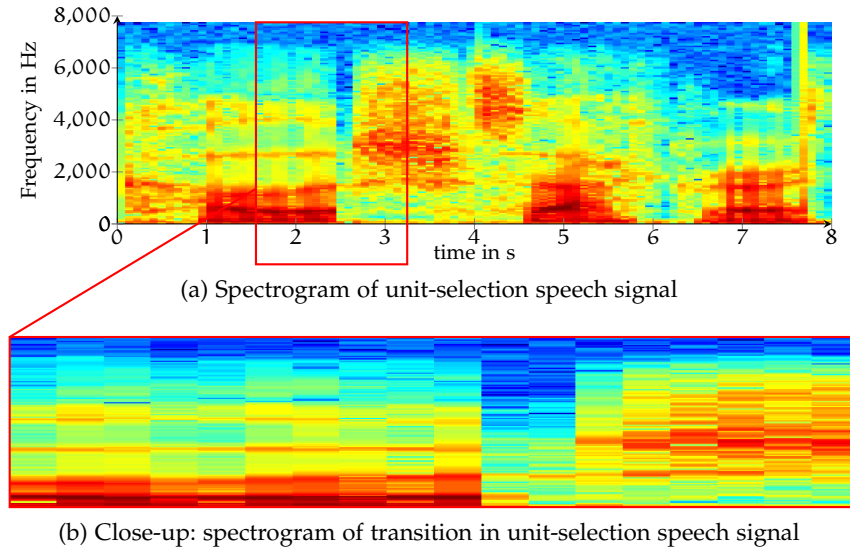


Figure 7.5: Spectrogram of a unit-selection speech signal and transitions.

7.1.2.2 Fourier-Based Features

In contrast to the result of a short-time Fourier transform (STFT), as visualized in Fig. 7.5, the Fourier transform omits any time information. Thus, a Fourier-based feature vector is proposed as (in the time domain) sudden changes of non-natural transitions cause (in frequency domain) higher amplitudes for higher frequencies. The resulting vector of the Fourier transform represents the amplitude as real part a and the phase as imaginary part $b i$. The magnitude of the signal $|a + b i|$ is calculated as $|a + b i| = \sqrt{a^2 + b^2}$.

7.1.2.3 Wavelet-Based Features

As depicted in Figs. 7.4 and 7.5, higher frequencies are more significant for distinguishing *bona fide presentations* from unit-selection samples than lower frequencies. A successive decomposition of a signal into bandpass-signals without information loss is possible, according to the Mallat theorem [271]. The discrete wavelet transform (DWT) can be understood as a bandpass filter decomposing the signal in iterative steps.

Earlier iterations provide higher frequency bands, later iterations provide lower ones. Assuming the discriminability of higher frequencies, a feature vector extracting the fifth detail level is examined. This choice was elaborated based on experimental results employing 10 343 *bona fide* and 10 461 attack samples. In order to cover multiple frequency bands establishing more discriminative robustness, the proposed DTW feature comprises information fused from third to fifth iteration. As the DWT represents a bandpass filter, the dimension of the result depends on the length of the analyzed signal. In order to obtain features with a fix dimension, a Fourier transformation is applied.

7.1.3 Analysis: Detection of Unit-Selection Presentation Attacks

The unit-selection attacks on the GSDC dataset are generated using Mary-TTS [265]. The proposed features are examined using support vector machines (SVM) and *Gaussian mixture model (GMM)* classifiers trained on the development set and optimized on the calibration set. A third partition of the GSDC is held back to validate the calibration results. The final performance of the classifiers is examined on the ASVspoof 2017 evaluation set, namely on the unit-selection partition (see chapter 3). In this section, the EER of *attack presentation classification error rate (APCER)* and *bona fide presentation classification error rate (BPCER)* is pointed out. The machine learning algorithms examined in this work are SVMs and GMMs. SVMs are chosen as they represent a well-established machine learning algorithm which provides binary classification and are known for good pattern recognition performance [268]. Following the assumption that Fourier based feature spaces comprise linear separable populations, linear SVM kernels may yield adequate performance. As an alternative to the SVM approach, GMMs are trained. *LLR* scores are computed on two 16-component GMMs, each representing *bona fide* and unit-selection speech. It is assumed that the proposed feature space results in different probabilistic clusters for each class.

Table 7.1: Configuration for best observed EER (on German speech data).

Feature	Classifier	EER	Number of Frequencies
DWT-fusion+FFT	SVM	5.0%	600
	GMM	5.6%	200
FFT	SVM	6.1%	1000
	GMM	6.3%	1100
DWT-5+FFT	SVM	23.1%	100
	GMM	20.0%	1600

7.1.3.1 Results on Calibration Set

The EER of the machine learning algorithms strongly depends on the size of the analyzed feature vector. In an analysis of the frequency resolutions from 100 to 3 000 (in steps of 100), the most promising configurations are investigated, cf. Tab. 7.1.

In general, a frequency resolution above 1100 bins leads to a rapidly increasing EER. This effect is likely to be caused by the machine learning algorithms as larger feature vectors require more training data in order to converge to satisfactory results.

For SVMs, the Fourier-based feature yields an EER of 6.1% with an FFT analyzing 1 000 frequencies. A feature fusion of FFT with the fifth iteration of a wavelet fusion (referred to as DWT-5+FFT) yields an EER of 23.1%. A feature that fuses the third to fifth DWT iteration features (referred to as DWT-fusion+FFT) exceeds the basic Fourier approach by 1.1 percent points, achieving 5.0%.

7.1.3.2 Results on Validation Set

The most promising configurations are examined on the validation set (German speech data). The observed EERs are depicted in Tab. 7.2. In general, the observed EERs are higher than those on the calibration set. On the validation dataset, the performance of the SVMs is less affected than the performance of the GMMs. The SVM classifying DWT-fusion+FFT features result in an EER of 7.1%, a drop of 2.1 percent points from the calibration set.

Fig. 7.6 depicts the **detection error trade-off (DET)** and **binary decision error trade-off (BET)** plots the examined algorithms. The performance DWT-5+FFT feature is behind the other features in all (visualized) operating points of interest. Assessed with the SVM, the DWT-5+FFT feature excels all other approaches for an APCER below 3%. The performances of the FFT and DWT countermeasures FFT+SVM and DWT-fusion+FFT+SVM are approximately identical in most operation points, FFT+GMM is slightly inferior.

Table 7.2: Best observed configurations evaluated with validation and ASVspoof sets.

Feature	Comparator	EER Validation set	EER ASVspoof set
DWT-fusion+FFT	SVM	7.1%	11.7%
	GMM	15.0%	24.6%
FFT	SVM	8.5%	22.6%
	GMM	9.5%	27.7%
DWT-5+FFT	SVM	27.0%	11.7%
	GMM	40.1%	45.7%
CFCCIF [266]	GMM	—	8.5%
CQCC [144]	GMM	—	0.5%

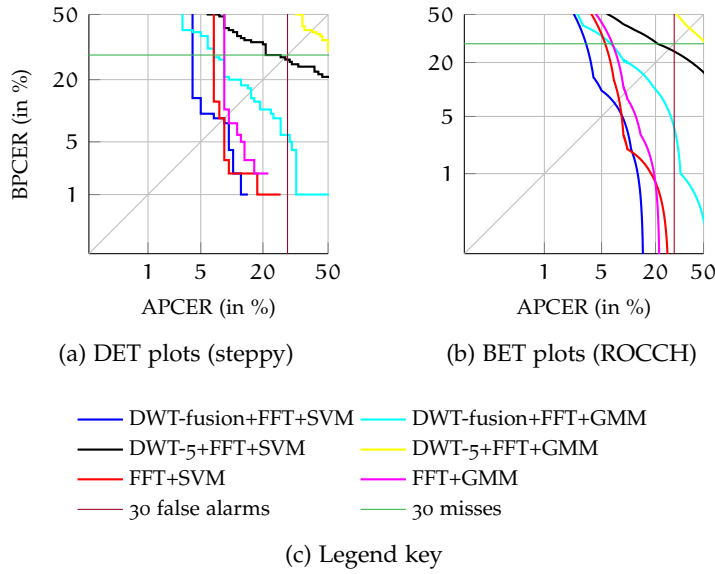


Figure 7.6: DET and BET plots for configurations on validation set.

7.1.3.3 Results on ASVspoof 2015 Unit-Selection Partition

Tab. 7.2 compares the performance on the ASVspoof 2015 unit-selection partition (English speech data) to the validation set (German speech data). The performance of the proposed algorithms (slightly) reduces on the validation set but changes (vastly) on the ASVspoof set. The EER of the DWT-fusion+FFT feature with SVM drops by 4.6 percent points to 11.7%. Remarkable is the performance increase of the analyzed DWT-5+FFT feature, improving the EER by 15.3 percent points to 11.7%. The error trade-off characteristic of the DWT-5+FFT feature with SVM yields the best observed DET and BET characteristics on the ASVspoof data, as depicted in Fig. 7.7.

The proposed features are able to detect unit-selection attacks with an EER of 7.1% on the GSDC and 11.7% on the ASVspoof unit-se-

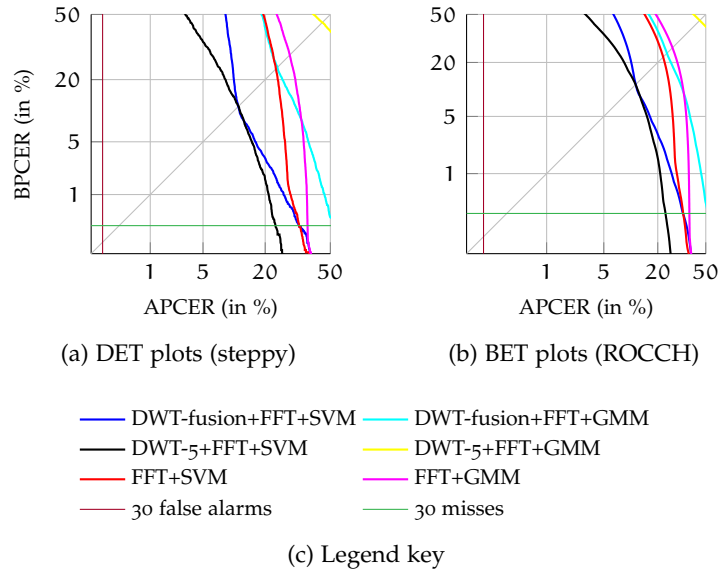


Figure 7.7: DET and BET plots for configurations with best EER on the ASVspoof unit-selection attacks.

lection set. Tab. 7.2 compares the proposed features and comparators to the best performing algorithm submitted at ASVspoof 2015 [266] and to the algorithm proposed in concurrence to this work [144]. To the best performing algorithm of the ASVspoof 2015 challenge, the introduced features DWT-5+FFT (SVM) and DWT-fusion+FFT (SVM) yield competitive results. The benefit of the proposed feature is that the computational costs are comparatively low since a Fourier transformation (FFT) is utilized instead of the more expensive spectrogram (STFT). The proposed feature space and classifiers represent a contrastive PAD system; in the presented approach, the unit-selection attack scheme to face is known during the approach development, which (on the contrary) is unknown for the countermeasures depicted in related work. A fusion of the above-mentioned features with low-level frequency analyses seems promising. Notably, this analysis comprised data shifts in terms of capture environments, the experimental setup, and the examined language. The field of voice PAD is actively researched in many international collaborations, especially the detection of replays, voice synthesis, and voice conversion, to countermeasure the various and also newly emerging forms of “fake audios” not only in voice biometrics but also in a multimedia engaging society.

7.2 PRIVACY AND DATA PROTECTION: HOMOMORPHICALLY ENCRYPTED BIOMETRIC INFORMATION AND COMPARATORS

The 2016 EU general data privacy regulation [272] declares biometric information as *personal data*, i.e., highly sensitive and entitled to the right of privacy preservation. Similarly, the current payment service directive [253] also requires biometric information protection to be employed in biometric systems utilized for banking services. To that end, the ISO/IEC IS 24745 [44] on biometric information protection provides guidance on how to preserve the *capture subject's* privacy. It defines the following three main properties to be fulfilled by protected biometric information:

- **Unlinkability.** Given only protected biometric information, it is not possible to say whether two protected biometric sample representations belong to the same subject. This prevents cross-comparisons for databases of different applications and ensures the privacy of the subject.
- **Renewability.** If a protected biometric reference is leaked or lost, the reference data can be revoked and renewed from the same biometric trait without the need to re-enroll.
- **Irreversibility.** Recovering biometric data from leaked protected biometric information is impossible without knowing the secret (key or algorithm) used to protect the biometric information. Restoring of valid biometric features or samples is thereby prevented.

In addition to these properties, other performance metrics, such as recognition accuracy, should be preserved.¹³⁴

Even if some authors argue that there is no need for biometric information protection (depending on the feature extraction) [274], sensitive information can be derived from unprotected biometric references, as has already been proved for other biometric characteristics [275, 276]. In particular, linkability of state-of-the-art speaker

¹³⁴ The implementation of privacy safeguards faces different challenges in the public and commercial sector. For standalone systems (commercials), the proposed biometric information protection scheme can be easily implemented on top of existing systems (which is not always the case, especially when following the privacy-by-design principle). For multi-owner systems (public sectors) that need to provide interoperability among different owners (forensic labs in different countries), any standalone solution for biometric information protection is difficult to implement. Ideally, interoperability is provided through standards which are implemented and tested for conformance. Nonetheless, the standardized biometric data interchange format—for forensic biometrics, the ANSI/NIST-ITL standard [273]—has been modified by almost every country (as of communication with forensic experts). This lack of a common version (one might suggest the Interpol version) makes the implementation of a common cryptosystem that still provides the demanded degree of freedom in interoperability harder. This discussion, however, lies outside the scope of this dissertation.

recognition features is demonstrated in [143] with the motivation of interchanging features among different voice biometric services. The interchange of biometric data across services is ethically addressed in [277], especially when targeting forensic scenarios. Accounting for latest data privacy legislation, biometric information protection is elaborated on in this work, especially for commercial but also for forensic (dual-use case) application scenarios, to demonstrate that LLR scores are computable with sustained precision in a distributed system architecture, while privacy is preserved and data is protected.

In this section¹³⁵, Paillier homomorphic encryption is made available to speaker recognition, targeting data privacy for subjects and *biometric system vendors*, investigating on the following hypotheses:

- Homomorphic encryption can be made available to generative comparators, considering the *two covariance model (2Cov)* comparator as a prototype scheme for *probabilistic linear discriminant analysis (PLDA)* comparators, sustaining the data privacy of subjects.
- Homomorphic encryption can be employed to protect biometric information as well as comparison model parameters.

7.2.1 Outline: Biometric Information Protection for Speaker Recognition

Current approaches to biometric information protection can be broadly classified into three categories [279]: i) cancelable biometrics [280], where irreversible transformations are applied at sample or feature level; ii) cryptobiometric systems [281], where a key is either bound or extracted from the biometric data; and iii) biometrics in the encrypted domain [282], where techniques based on *homomorphic encryption (HE)* and garbled circuits are used to protect the data. While cancelable biometrics and cryptobiometric systems usually report some accuracy degradation [279], the use of HE schemes prevents such loss, since the operations carried out in the encrypted domain are equivalent to those performed with plaintext data. For this reason, HE schemes in this work are applied similarly to the ones proposed in [283–286] to speaker recognition relying on generative comparators, such as PLDA. Here, data privacy is ensured for data capture subjects as well as for comparison models using the 2Cov approach [133, 158] (which is the full subspace PLDA model; here as a prototype generative comparison algorithm).

¹³⁵ Parts of this section are based on a collaborative work with Sergey Isadskiy, Jascha Kolberg, Marta Gomez-Barrero, and Christoph Busch [74] (*Odyssey 2018 best paper award*), which emerged from the collaboration with Sergey on his master's thesis [278], as well as the work with Abelino Jimenez, Amos Treiber, Jascha Kolberg, and Nicholas Evans (among others) [19], a 23 co-author collaboration survey on data privacy in speaker and speech characterization.

Comparison scores of generative and discriminative models can be probabilistic. In contrast to discriminative comparators, generative models can emit features with associated likelihoods based on pre-trained models and thus estimate probabilistic similarity. The parameters of pre-trained models deserve protection by the biometric system vendors who distribute the sensitive parameters to various biometric service operators. A mutual encryption scheme granting subject and vendor data privacy is further proposed by employing well-established Paillier homomorphic cryptosystems [287, 288], blindfolding operators. Notably, while conventional image based biometric systems employ comparators operating either on binary or nonnegative integers [285, 286, 289, 290], the generative comparators used in speaker recognition applications make assumptions on underlying distributions, such as normal distribution [127, 130], consequently operating on normally distributed float values.

7.2.2 Overview: Biometric Information Protection

In order to apply biometric information protection schemes of standardization, binarization¹³⁶ can be employed [44]. Related work on the binarization of traditional speaker recognition systems utilizing [universal background models \(UBMs\)](#) targeting the GMM–UBM approach can be found in [291–293]. In addition, [294] proposes a biometric information protection scheme for speaker recognition, based on binarized GMM supervectors.

Due to the binarization process, however, the biometric performance usually declines, and calibration properties are lost. Contrary to performance-lossy information protection approaches such as biometric cryptosystems and cancelable biometrics [279], HE completely preserves biometric accuracy. Therefore, Paillier HE schemes are investigated here, which are already introduced to other biometric modalities, such as face [295], signature [283], iris [289], and fingerprint [290] recognition, considering Hamming distances (XOR operator), dynamic time warping (DTW), the Euclidean distance, and the cosine similarity. Thus, for the remainder of this section, the focus is put on homomorphic cryptosystems.

In [296] and [297], the authors provide an overview of several biometric information protection schemes based on HE and garbled circuits. Barni et al. [290] present a way to protect fixed-length finger-codes [298] using HE. This system is modified in [299] in order to accelerate the process by reducing the size of the fingercode. However, a reduction of information also leads to a degradation of biometric recognition performance. Ye et al. present an anonymous biometric

¹³⁶ Although it might appear counterintuitive to (re-) binarize data which is already stored in a binary format (floats are discrete data after all), the term *binarization* is often referred to in the literature.

access control (ABAC) system [300] for iris recognition. Their system setup only verifies whether a subject is enrolled without revealing her identity and thus grants anonymity towards the subject. Another ABAC protocol is proposed in [301] by Luo et al., and a secure similarity search algorithm is presented for anonymous authentication. Combining homomorphic encryption with garbled circuits, Blanton and Gasti [302] implement secure protocols for iris and fingerprint recognition.

Among the existing cryptosystems in the literature, the encryption algorithms based on lattices are assumed to be post-quantum secure [303, conjecture 2], which is a convenient property for a public key encryption scheme. Using ideal lattices in a somewhat homomorphic encryption (SHE) scheme, Yasuda et al. [304] efficiently compute the Hamming distance of encrypted references and probes by using a packing method before the encryption. By using binary feature vectors with a constant size of 2048 bits for all biometric data, Yasuda et al. [305] present a new packing method in a SHE scheme for biometric authentication based on a special version of the ring learning with errors assumption. Another privacy-preserving biometric authentication approach [306] splits a 2048 bits iris code into 64 blocks with 32 bits each and encrypts these blocks using n -th degree truncated polynomial ring (NTRU). As in the aforementioned works, scores are computed in the encrypted domain without disclosure of biometric information.

HE and garbled circuits schemes are not new to the speech communication community (including speaker recognition). They have been studied by the research group of Bhiksha Raj, cf. [307–311] (among other works) for almost a decade. Regarding HE, their work considers mostly protocols and cryptosystems for hidden Markov models (HMMs) and GMMs. This dissertation contributes HE for PLDA/2Cov comparison of probabilistic and discriminative embeddings, i.e., for the state-of-the-art comparison of *i-vectors* and *x-vectors*.

7.2.3 Homomorphic Cryptosystems

Homomorphic encryption (HE) [312–314] has the property that computations on the ciphertext are equivalent to those carried out on the plaintext. Homomorphisms are functions which preserve algebraic structures of groups [315]. The function $f : G \rightarrow H$ is a homomorphism for the two groups $(G, \diamond), (H, \square)$ with sets G, H and operators \diamond, \square if:

$$f(g \diamond g') = f(g) \square f(g') \quad \forall g, g' \in G. \quad (7.1)$$

Example: Homomorphisms for Public-Key Cryptography

This property is used for public-key cryptosystems. Thereby, the function f is seen as an encryption function using a public key. The plaintext operation \diamond is the biometric comparison (e.g., PLDA); its result is a (plaintext) score. Its equivalent for ciphertexts is the operation \square , i.e., biometric comparison for encrypted references and encrypted probes ($f(g) \square f(g')$); its result is equivalent to an encrypted score $f(g \diamond g')$ (the homomorphism). The inverse of f , the function f^{-1} , is the decryption which uses the secret key, such that scores computed on ciphertexts can be decrypted.

In distributed system architectures, privacy is preserved, whereas an authentication server alone knows the secret key. The secret key's security needs to be guaranteed. By contrast, the public key is known by everyone. Client devices use the public key to encrypt reference and probe features. For privacy, it is of utmost importance that the entities involved with the authentication server (those who could access the secret key and encrypt protected references and probes) are prohibited from accessing protected data but encrypted scores.

In the case of a database leakage with encrypted references, no one but the holders of the secret key are capable of decryption (the assumption is that without knowing the secret key, decryption is infeasible and takes too much computational time). The responsibility of the secret key holders is then to generate a new public/secret key pair with which the protected database is *renewed*: an authority is granted access to the old secret key and the new public key to decrypt (with the old secret key) and encrypt (with the new public key) the biometric data. This property (*renewability*) is, beside the *unlinkability* and *irreversibility* properties within biometric information protection, crucial.

Public-key cryptosystems $(K, M, C, \text{enc}, \text{dec})$ with sets of keys K , plaintexts M , ciphertexts C , and functions representing encryption enc and decryption dec are homomorphic if:

$$\begin{aligned} &\forall m_1, m_2 \in M, \forall pk \in K : \\ &\text{enc}_{pk}(m_1) \square \text{enc}_{pk}(m_2) = \text{enc}_{pk}(m_1 \diamond m_2), \end{aligned} \quad (7.2)$$

where the public key pk is used for encryption and the secret key sk for the decryption functions, respectively:

$$\begin{aligned} &\text{enc}_{pk} : M \rightarrow C, \\ &\text{dec}_{sk} : C \rightarrow M. \end{aligned} \quad (7.3)$$

7.2.3.1 Paillier HE Scheme

Motivated by asymmetric Paillier cryptosystems [287, 288], HE has been made available to biometric information protection [283, 285,

286]. Paillier cryptosystems are homomorphic, probabilistic encryption schemes based on the decisional composite residuosity assumption (DCRA) [287]: *for integers n, z it is hard to decide whether z is an n -residue modulo n^2* .¹³⁷ Due to this assumption, the Paillier cryptosystem is secure against *honest but curious users* conducting chosen ciphertext attacks [287, 316, 317].

In the Paillier cryptosystem, the public key $pk = (n, g)$ is defined by $n = p \cdot q$ and $g \in \mathbb{Z}_{n^2}^*$, where p, q are two large prime numbers, such that $\gcd(p \cdot q, (p-1)(q-1)) = 1$, and with $\mathbb{Z}_{n^2}^*$ as the set of module n^2 integers having a modular multiplicative inverse. Based on p, q , the secret key $sk = (\lambda, \mu)$ is defined by $\lambda = \text{lcm}(p-1, q-1)$ and $\mu = \bar{\rho} \mod n$. Thereby, $\bar{\rho}$ is the modular multiplicative inverse to $\rho = L(g^\lambda \mod n^2)$, where $L(x) = \frac{x-1}{n}$. The modular multiplicative inverse $\bar{\rho}$ to ρ is defined as $\rho \bar{\rho} \equiv 1 \pmod{n^2}$. In modulo n^2 , the multiplication of ρ and its inverse results in the identity (the *mathematical identity*, an equality relation, such as the value 1 for ordinary multiplication). By consequence, both terms (ρ and $\bar{\rho}$) are relatively prime (coprime) to another: $\gcd(\rho, \bar{\rho}) = 1$.

During encryption $c = \text{enc}_{pk}(m, s) \in \mathbb{Z}_{n^2}^*$ of a message $m \in \mathbb{Z}_n$ with public key pk , a random number $s \in \mathbb{Z}_n^*$ provides the probabilistic nature of the cryptosystem, i.e., $\text{enc}_{pk}(m, s_1) \neq \text{enc}_{pk}(m, s_2)$ for two different $s_1, s_2 \in \mathbb{Z}_n^*$:

$$c = \text{enc}_{pk}(m, s) = g^m s^n \mod n^2, \quad (7.4)$$

which is abbreviated as $\text{enc}_{pk}(m)$ in the following. Ciphertexts are decrypted as:

$$m = \text{dec}_{sk}(c) = L(c^\lambda \mod n^2) \mu \mod n. \quad (7.5)$$

Similarly to [283, 285, 286, 288, 307–309], the following additive homomorphic properties of the Paillier cryptosystem are used regarding plaintexts m_1, m_2 and corresponding ciphertexts c_1, c_2 :

$$\begin{aligned} \text{dec}_{sk}(c_1 c_2) &= m_1 + m_2 \mod n, \\ \text{dec}_{sk}(c_1^l) &= m_1 l \mod n \quad \text{with a constant } l. \end{aligned} \quad (7.6)$$

In other words, while the decrypted product of two ciphertexts is equivalent to the sum of two plaintexts, the decrypted exponentiation of a ciphertext and a (plaintext) constant l is equivalent to the product of the corresponding plaintext m_1 and this (plaintext) constant.

7.2.3.2 Homomorphic Biometric Information Protection

Targeting biometric information protection, data privacy friendly comparison schemes are sought in which only encrypted references

¹³⁷ In other words, while it is easy to multiply integers y n -times with another and to compute the remainder z after division by n^2 , i.e., $z \equiv y^n \pmod{n^2}$, it is assumed to be hard to find the existence (and value) of y for given numbers z, n .

(no plaintext references) are stored in databases. As such, the Euclidean and cosine similarity comparison scores S_{Euc}, S_{cos} between two D -dimensional vectors $\mathbf{X} = \{x_1, \dots, x_D\}, \mathbf{Y} = \{y_1, \dots, y_D\}$ are computationally derived as [283, 285, 286]:

$$S_{Euc}(\mathbf{X}, \mathbf{Y}) = \sum_{d=1}^D x_d^2 + \sum_{d=1}^D y_d^2 - 2 \sum_{d=1}^D x_d y_d, \quad (7.7)$$

and the corresponding encrypted score $\text{enc}_{pk}(S_{Euc}(\mathbf{X}, \mathbf{Y}))$:

$$\begin{aligned} & \text{enc}_{pk}(S_{Euc}(\mathbf{X}, \mathbf{Y})) = \\ & \text{enc}_{pk}\left(\sum_{d=1}^D x_d^2\right) \text{enc}_{pk}\left(\sum_{d=1}^D y_d^2\right) \prod_{d=1}^D \text{enc}_{pk}(y_d)^{-2x_d}, \end{aligned} \quad (7.8)$$

where the protected reference $\mathbf{Y}_{Euc}^{\text{enc}_{pk}}$ is defined as:

$$\mathbf{Y}_{Euc}^{\text{enc}_{pk}} = \left(\text{enc}_{pk}\left(\sum_{d=1}^D y_d^2\right), (\text{enc}_{pk}(y_d))_{d=1}^D \right). \quad (7.9)$$

On the other hand, the cosine comparison is derived as [283, 285]:

$$\begin{aligned} S_{cos}(\mathbf{X}, \mathbf{Y}) &= \frac{\mathbf{X}^T \mathbf{Y}}{\|\mathbf{X}\| \|\mathbf{Y}\|} = \sum_{d=1}^D \frac{x_d y_d}{\|\mathbf{X}\| \|\mathbf{Y}\|}, \\ \text{enc}_{pk}(S_{cos}(\mathbf{X}, \mathbf{Y})) &= \prod_{d=1}^D \text{enc}_{pk}\left(\frac{y_d}{\|\mathbf{Y}\|}\right)^{\frac{x_d}{\|\mathbf{X}\|}}, \end{aligned} \quad (7.10)$$

where the protected reference $\mathbf{Y}_{cos}^{\text{enc}_{pk}}$ is defined for length-normalized features as:

$$\mathbf{Y}_{cos}^{\text{enc}_{pk}} = \left((\text{enc}_{pk}(y_d))_{d=1}^D \right) = \text{enc}_{pk}(\mathbf{Y}). \quad (7.11)$$

In [283, 285], solely positive integers are considered. Accommodating a broader range of positive only float values, a 10^{12} scaling factor is employed. Accounting for negative float values, this study relies on an alternative float representation.

Fig. 7.8 illustrates a distributed client-server architecture employing HE with a cosine comparison: a client device C extracts the probe feature vector \mathbf{X} and requests the encrypted reference feature vector $\text{enc}_{pk}(\mathbf{Y})$ from the database $\text{DB}_{controller}$ (which is in the province of the data [biometric data controller](#)). Scores are then calculated on the client device (here as the [biometric data processor](#)) and sent to the authentication server $\text{AS}_{operator}$ (in the province of, e.g., [biometric system owners, operators and providers](#)), which holds the key pair (pk, sk) . Based on a pre-defined threshold, $\text{AS}_{operator}$ outputs the decision of whether or not the decrypted score S_{cos} is greater or equal to a threshold η .¹³⁸ Ideally, the $\text{DB}_{controller}$ is in the domain of an independent

¹³⁸ Throughout this section, the [LLR](#) threshold notation η is used to indicate that scores are assumed to be well-calibrated before threshold comparison.

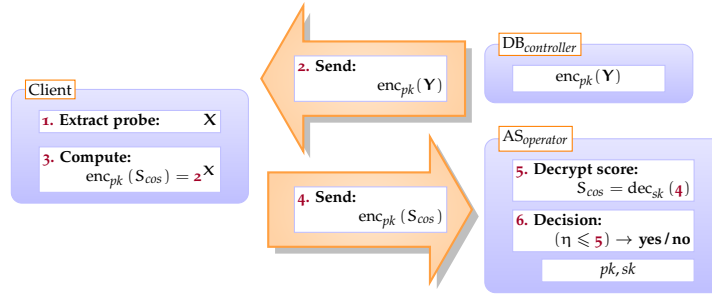


Figure 7.8: Architecture of homomorphically encrypted cosine similarity comparison for length-normalized features, cf. [283], with client device, servers (blue boxes) and communication channels (orange arrows).

Table 7.3: Complexity analysis for the Euclidean and cosine comparators during verification, cf. [283], assuming $D = 250$ dimensional features, the size of an encrypted feature $v = 0.5$ KiB, and the plain feature size $p = 64$ bits.

	Euclidean	Cosine
N° encryptions	D	0
N° decryptions	1	1
N° additions	$D - 1$	0
N° products	$2D + 4$	$D - 1$
N° exponentiations	$2D$	D
Plain reference size	pD ≈ 2.0 KiB	pD ≈ 2.0 KiB
Protected reference size	$v(D + 1)$ $= 125.5$ KiB	vD $= 125.0$ KiB
Channels: amount of protected data exchanged	$v(D + 2)$ $= 126.0$ KiB	$v(D + 1)$ $= 125.5$ KiB

data controller, restricting access to operators, among others. Tab. 7.3 provides an overview of the complexity of the encrypted Euclidean and cosine comparison. Numbers diverge from [283] as in the signature recognition scenario, five references are encrypted rather than, e.g., an averaged reference model. As references are encrypted during enrolment, cosine-based biometric comparisons require no additional encryptions, whereas in Euclidean-based comparisons, the probe features need to be encrypted.

7.2.4 Revisiting PLDA and 2Cov Comparators in Speaker Recognition

State-of-the-art **i-vector** comparators belong to the **PLDA** family [157, 158]. PLDA comparators conduct a likelihood ratio scoring comparing the probabilities of the propositions that reference and probe i-vectors X, Y stemming from \mathcal{A} : the same source or from \mathcal{B} : differ-

ent sources. Therefore, within and between speaker variabilities are examined in a latent (biometric) feature subspace. In this work, emphasis is put on the 2Cov approach [133, 158], the full-subspace Gaussian PLDA. Notably, the 2Cov comparator can also be related to pairwise support vector machines [133, 158]. For the sake of tractability, this study focuses on the generative 2Cov model. Also, i-vectors are solely considered as point estimates, assuming ideal precision during feature extraction. The closed-form solution to the 2Cov scoring is denoted regarding within and between covariances $\mathbf{W}^{-1}, \mathbf{B}^{-1}$ with mean $\boldsymbol{\mu}$ [133], see Eq. 2.93:

$$\begin{aligned}
 S_{2Cov}(\mathbf{X}, \mathbf{Y}) &= \mathbf{X}^T \boldsymbol{\Lambda} \mathbf{Y} + \mathbf{Y}^T \boldsymbol{\Lambda} \mathbf{X} + \mathbf{X}^T \boldsymbol{\Gamma} \mathbf{X} + \mathbf{Y}^T \boldsymbol{\Gamma} \mathbf{Y} + \mathbf{c}^T (\mathbf{X} + \mathbf{Y}) + k \\
 \text{with } \boldsymbol{\Lambda} &= \frac{1}{2} \mathbf{W}^T \tilde{\boldsymbol{\Lambda}} \mathbf{W}, \quad \boldsymbol{\Gamma} = \frac{1}{2} \mathbf{W}^T (\tilde{\boldsymbol{\Lambda}} - \tilde{\boldsymbol{\Gamma}}) \mathbf{W}, \\
 \mathbf{c} &= \mathbf{W}^T (\tilde{\boldsymbol{\Lambda}} - \tilde{\boldsymbol{\Gamma}}) \mathbf{B} \boldsymbol{\mu}, \quad k = \tilde{k} + \frac{1}{2} \left((\mathbf{B} \boldsymbol{\mu})^T (\tilde{\boldsymbol{\Lambda}} - 2\tilde{\boldsymbol{\Gamma}}) \mathbf{B} \boldsymbol{\mu} \right), \\
 \tilde{\boldsymbol{\Lambda}} &= (\mathbf{B} + 2\mathbf{W})^{-1}, \quad \tilde{\boldsymbol{\Gamma}} = (\mathbf{B} + \mathbf{W})^{-1}, \\
 \tilde{k} &= 2 \log |\tilde{\boldsymbol{\Gamma}}| - \log |\tilde{\boldsymbol{\Lambda}}| - \log |\mathbf{B}| + \boldsymbol{\mu}^T \mathbf{B} \boldsymbol{\mu}. \tag{7.12}
 \end{aligned}$$

7.2.5 Contribution: Privacy Architectures

In the following section, two discriminative HE schemes are proposed. The first emphasizes HE for (i-vector) embeddings during 2Cov comparison, seeking data privacy for *biometric capture subjects*, whereas the second focuses on the encryption of embeddings as well as 2Cov model parameters, targeting data protection for subjects and vendors. An auxiliary float representation is implemented, encoding float values as nonnegative integers for the purpose of providing Paillier properties, cf. Eq. (7.6).

7.2.5.1 Auxiliary Float Representation: Nonnegative Integers

For the purpose of representing float values of i-vector embeddings as nonnegative integer values, i.e., seeking conformance to Paillier cryptosystems, the integer encoding scheme standardized in IEEE 754 is employed [318]. Floats are encoded by four terms S, M, B, E as $S \times M \times B^E$ (similar to scientific number notation): a boolean flag S represents the sign of the float, an unsigned integer M represents the *mantissa* of the float, an unsigned integer $B = 16$ represents the base of the float, and an unsigned integer E represents the exponent of the float. Non-negative integers are derived by seeking congruent positive representations in modulo n^2 , i.e., regarding the public key domain. Accounting for negative values [319], the plaintext integer domain is divided into four intervals: $[0, \frac{n}{3})$ for positive float representations, $[\frac{2n}{3}, n)$ for negative float representations, and $[\frac{n}{3}, \frac{2n}{3})$ as well as $[n, \infty)$ for the purpose of detecting overflows resulting from previous Paillier HE operations. Targeting Paillier HE, the same exponents

of m_1, m_2 are required. The mantissa is hence encrypted as a nonnegative integer representation. The plaintext exponent of the depending mantissa encoding is kept auxiliary.¹³⁹ Security requirements are met due to the DCRA employing randomized mantissa obfuscation during encryption. In Paillier addition, encrypted mantissae are scaled for equivalent addend exponents. In Paillier multiplication, modular exponentiation of $c = \text{enc}_{pk}(M, s)$ is conducted, during which mantissae are kept rather small by iterative multiplications than by right-away exponentiation.

7.2.5.2 Data Privacy: Protecting Subjects

For the sake of tractability, a zero mean is assumed, causing $c = 0$, and neglecting the normalization term, i.e., $k = 0$, such that the following scheme solely holds for discriminative 2Cov . However, calibrated scores can be easily achieved by adding the k term after score decryption:

$$\begin{aligned} S_{2\text{Cov}}(\mathbf{X}, \mathbf{Y}) &= \mathbf{X}^T \mathbf{\Lambda} \mathbf{Y} + \mathbf{Y}^T \mathbf{\Lambda} \mathbf{X} + \mathbf{X}^T \mathbf{\Gamma} \mathbf{X} + \mathbf{Y}^T \mathbf{\Gamma} \mathbf{Y} \\ &= (\mathbf{X}^T \mathbf{\Lambda}) \mathbf{Y} + \mathbf{Y}^T (\mathbf{\Lambda} \mathbf{X}) + \mathbf{X}^T \mathbf{\Gamma} \mathbf{X} + \mathbf{Y}^T \mathbf{\Gamma} \mathbf{Y}. \end{aligned} \quad (7.13)$$

For the discriminative 2Cov , HE is employed, motivated by the (symmetric) dot product for vector multiplication:

$$\begin{aligned} \text{enc}_{pk}(\mathbf{Y})^{\mathbf{X}} &= \prod_{d=1}^D \text{enc}_{pk}(y_d)^{x_d} = \text{enc}_{pk}(\mathbf{X}^T \mathbf{Y}) \\ &= \text{enc}_{pk}(\mathbf{Y}^T \mathbf{X}) = \prod_{d=1}^D \text{enc}_{pk}(x_d)^{y_d} = \text{enc}_{pk}(\mathbf{X})^{\mathbf{Y}}, \\ \text{enc}_{pk}(S_{2\text{Cov}}(\mathbf{X}, \mathbf{Y})) &= \text{enc}_{pk}(\mathbf{Y})^{(\mathbf{X}^T \mathbf{\Lambda})} \text{enc}_{pk}(\mathbf{Y})^{(\mathbf{\Lambda} \mathbf{X})} \\ &\quad \text{enc}_{pk}(\mathbf{X}^T \mathbf{\Gamma} \mathbf{X}) \text{enc}_{pk}(\mathbf{Y}^T \mathbf{\Gamma} \mathbf{Y}), \\ \text{enc}_{pk}(\mathbf{Y}) &= (\text{enc}_{pk}(y_d))_{d=1}^D \end{aligned} \quad (7.14)$$

with auxiliary vectors are denoted as $(\mathbf{X}^T \mathbf{\Lambda}), (\mathbf{\Lambda} \mathbf{X})$, and the protected reference $\mathbf{Y}_{2\text{Cov}}^{\text{enc}_{pk}} = (\text{enc}_{pk}(\mathbf{Y}), \text{enc}_{pk}(\mathbf{Y}^T \mathbf{\Gamma} \mathbf{Y}))$.

Fig. 7.9 illustrates the proposed HE architecture for a distributed system. Similar to the cosine comparison HE approach, the scores are computed in the encrypted domain on the client device and

¹³⁹ According to [319], by choosing a value for the base, the information leakage of the plaintext exponent is outlined. By contrast, when varying the base term, decoders need to be manually informed. When further denoting a fixed exponent, one needs to decide on the minimal precision lost: when base and exponent terms are fixed, float values are effectively multiplied with some large fixed integer, and remaining precision digits are truncated (lost). By finding a lower bound for naturally arising exponents in the data, a practical exponent term can be set. For plaintext exponents, smaller base terms exacerbate information leakage.

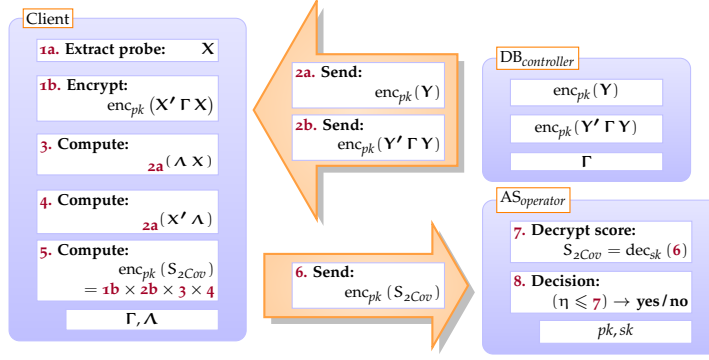


Figure 7.9: Contribution: architecture of homomorphically encrypted PLDA/2Cov comparison solely protecting subject data with client device, servers (blue boxes), and communication channels (orange arrows).

decrypted on the authentication server. Thereby, the 2Cov score is computed in four parts. Notably, Eq. (7.13) computationally equals Eq. (2.91), i.e., carried out 2Cov score computations are the same algorithmic computations for Gaussian PLDA (during recognition, not during training). Thus, the proposed architecture is also applicable and extensible to other members of the PLDA comparator family.

7.2.5.3 Privacy and Data Security: Protecting Subjects and Vendors

Contrary to established biometric HE approaches employing non-generative comparators, generative comparators require trained hyper-parameters, e.g., between and within covariance matrices in terms of the 2Cov comparator. For the purpose of protecting subject as well as vendor data, two key sets are employed $(pk_1, sk_1), (pk_2, sk_2)$. Using the Frobenius inner product¹⁴⁰, Eq. (7.12) is restated [133], see Eq. (2.94):

$$\begin{aligned}
 S_{2Cov}(X, Y) &= \langle \Lambda, XY^T + YX^T \rangle + \langle \Gamma, XX^T + YY^T \rangle + c^T(X + Y) + k, \\
 &= w_{\Lambda}^T \varphi_{\Lambda}(X, Y) + w_{\Gamma}^T \varphi_{\Gamma}(X, Y) \\
 &\quad + w_c^T \varphi_c(X, Y) + w_k^T \varphi_k(X, Y), \\
 &= w^T \varphi(X, Y) \quad \text{with} \\
 \varphi(X, Y) &= \begin{bmatrix} \text{vec}(XY^T + YX^T) \\ \text{vec}(XX^T + YY^T) \\ X + Y \\ 1 \end{bmatrix} = \begin{bmatrix} \varphi_{\Lambda}(X, Y) \\ \varphi_{\Gamma}(X, Y) \\ \varphi_c(X, Y) \\ \varphi_k(X, Y) \end{bmatrix},
 \end{aligned}$$

¹⁴⁰ The inner Frobenius product denotes $x^T A y = \langle A, x y^T \rangle = \text{vec}(A)^T \text{vec}(x y^T)$, where $\text{vec}(\cdot)$ denotes the operator stacking matrices into a vector and $\langle A, B \rangle$ is the dot product between matrices, cf. [133].

$$\mathbf{w} = \begin{bmatrix} \text{vec}(\mathbf{A}) \\ \text{vec}(\mathbf{B}) \\ \mathbf{c} \\ \mathbf{k} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_A \\ \mathbf{w}_B \\ \mathbf{w}_c \\ \mathbf{w}_k \end{bmatrix}. \quad (7.15)$$

For the 2Cov, a mutual HE scheme sustaining data privacy for subjects and vendors can be employed by extending the inner product of vectors to the Frobenius inner product of matrices \mathbf{A}, \mathbf{B} . The latter can be reformulated via the $\text{vec}(\cdot)$ operator as the inner product of (column-stacked) vectors, such that the dot product can be employed with a public key pk as well:

$$\text{enc}_{pk}(\mathbf{A})^{\langle \cdot \rangle(\mathbf{B})} = \text{enc}_{pk}(\text{vec}(\mathbf{A}))^{\text{vec}(\mathbf{B})}, \quad (7.16)$$

where the encryption of a matrix \mathbf{A} is denoted as:

$$\text{enc}_{pk}(\mathbf{A}) = \left((\text{enc}_{pk}(a_{i,j}))_{i,j=1}^D \right)_{j=1}^D. \quad (7.17)$$

In the simplified 2Cov comparator, the encrypted vendor and operator communication takes the form:

$$\begin{aligned} S_{2Cov}(\mathbf{X}, \mathbf{Y}) &= \mathbf{w}_A^T \varphi_A(\mathbf{X}, \mathbf{Y}) + \mathbf{w}_B^T \varphi_B(\mathbf{X}, \mathbf{Y}), \\ \text{enc}_{pk_2}(S_{2Cov}(\mathbf{X}, \mathbf{Y})) &= \text{enc}_{pk_2}(\mathbf{A})^{\langle \cdot \rangle(\mathbf{c}_1)} \text{enc}_{pk_2}(\mathbf{B})^{\langle \cdot \rangle(\mathbf{c}_2 + \mathbf{c}_3)} \\ \text{with } \mathbf{c}_1 &= \mathbf{X}\mathbf{Y}^T + \mathbf{Y}\mathbf{X}^T, \mathbf{c}_2 = \mathbf{X}\mathbf{X}^T, \mathbf{c}_3 = \mathbf{Y}\mathbf{Y}^T, \end{aligned} \quad (7.18)$$

where the computation of $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$ is subdue to the encrypted operator, controller, and client device communication:

$$\begin{aligned} \text{enc}_{pk_1}(\mathbf{c}_1) &= \text{enc}_{pk_1}(\mathbf{Y})^{\mathbf{X}^T} \circ \text{enc}_{pk_1}(\mathbf{Y}^T)^{\mathbf{X}}, \\ \text{enc}_{pk_1}(\mathbf{c}_2 + \mathbf{c}_3) &= \text{enc}_{pk_1}(\mathbf{X}\mathbf{X}^T) \circ \text{enc}_{pk_1}(\mathbf{Y}\mathbf{Y}^T). \end{aligned} \quad (7.19)$$

Here, \circ denotes the Hadamard product¹⁴¹, and the terms $\text{enc}_{pk_1}(\mathbf{Y})^{\mathbf{X}^T}$, $\text{enc}_{pk_1}(\mathbf{Y}^T)^{\mathbf{X}}$ represent exponentiations in an outer product fashion, resulting in the matrices $\text{enc}_{pk_1}(\mathbf{Y}\mathbf{X}^T)$ and $\text{enc}_{pk_1}(\mathbf{X}\mathbf{Y}^T)$, respectively. Finally, the protected reference is $\mathbf{Y}_{2Cov}^{\text{enc}_{pk_1}} = (\text{enc}_{pk_1}(\mathbf{Y}), \text{enc}_{pk_1}(\mathbf{Y}\mathbf{Y}^T))$.

Fig. 7.10 presents the proposed architecture. The previously proposed architecture is extended by additional communication channels between operators and vendors. Applications employ two key pairs, such that privacy preservation and data protection can be achieved dependent on both: i) different biometric services of an **biometric system operator, owner, or provider**, and ii) multiple provisions of a biometric system to service operators, owners, and providers by a **biometric system vendor**. Consequently, additional servers are necessary on the vendor site in terms of a database $\text{DB}_{\text{vendor}}$ and an authentication server $\text{AS}_{\text{vendor}}$. As the computations carried out during

¹⁴¹ The Hadamard product is an entrywise product of two matrices \mathbf{A}, \mathbf{B} with the same dimension: $\mathbf{A} \circ \mathbf{B} = (\mathbf{A})_{i,j} (\mathbf{B})_{i,j}$.

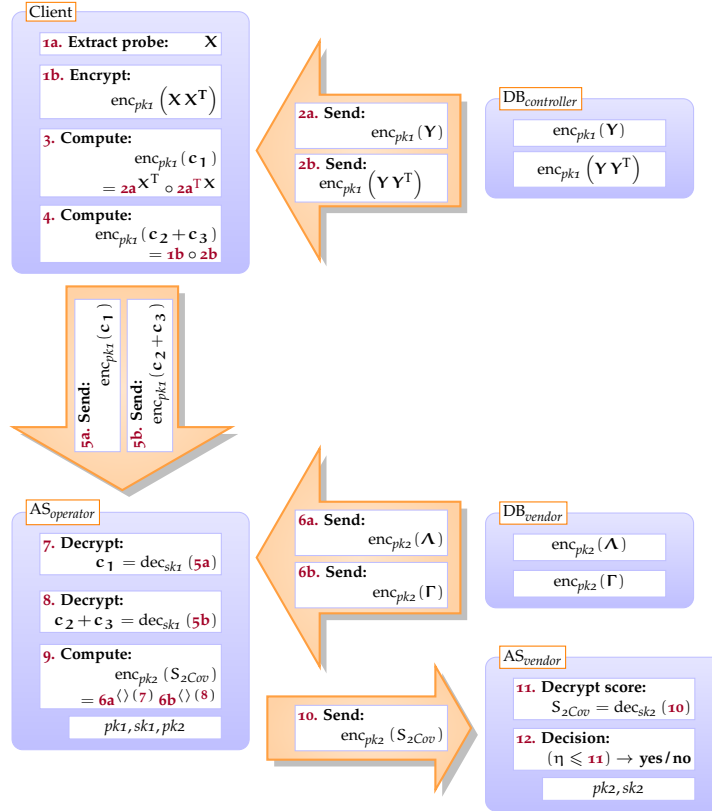


Figure 7.10: Contribution: architecture of protected biometric information and comparison model hyper-parameters with client device, servers (blue boxes), and communication channels (orange arrows).

verification are the same for the depicted 2Cov comparator and for other Gaussian PLDA comparators, the proposed architecture is also applicable and extensible to other members of the PLDA comparator family.

7.2.6 Proof-of-Concept Study

The prototype system comprises a dimension reduction to $D = 250$ by linear discriminant analysis, within class covariance normalization, length normalization, and 2Cov comparison. For the Paillier cryptosystem, $n = 2048$ bits keys are used in accordance with the [US National Institute of Standards and Technology \(NIST\)](#) recommendation SP 800-56A revision 3 [320]. In contrast, plaintext operations consider double floating-point precision, i.e., $p = 64$ bits per plain real feature value. Implementations are based on the freely available *sidekit* [146] and *Python Paillier* [319]. Fig. 7.11 illustrates the [DET](#) and [BET](#) performance of conventional and HE 2Cov comparators on the evaluation set in terms of [false non-match rate \(FNMR\)](#) and [false match rate \(FMR\)](#): the baseline performance is preserved across all

operating points (the same for all systems). The DET and BET are depicted in terms of the ROC's convex hull (ROCCH). For the exemplary 2Cov system, a 2.5% EER, a 0.050 minimum DCF (minDCF) (parameterized according to [36]), and a $0.099 C_{llr}^{\min}$ are preserved in the protected domain. As the k normalization term is neglected in this setup, the baseline system yielded a 9.560 C_{llr} . Calibration loss can be reduced by a post score re-bias or by employing conventional score calibration methods, cf. [79]. By using linear score calibration trained on the development set with known labels, C_{llr} is reduced to 0.284.

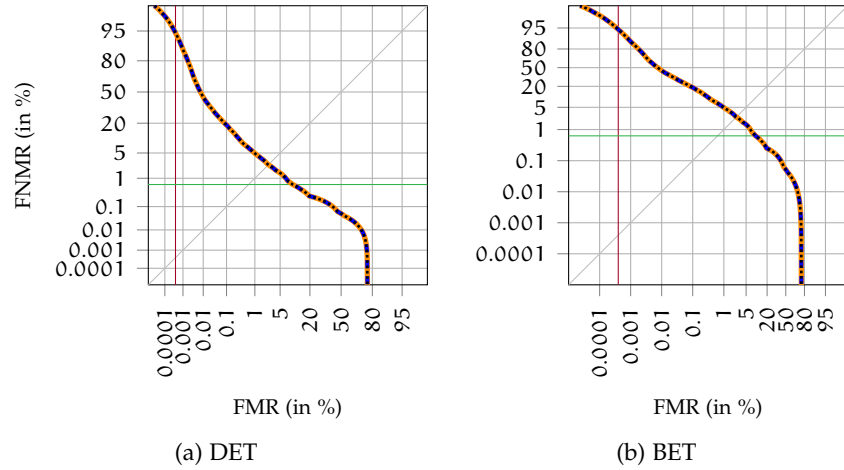


Figure 7.11: DET and BET comparison of the baseline PLDA/2Cov system (orange) and the proposed HE PLDA/2Cov schemes, focusing on subject data protection (blue, dashed) and the protection of subject and vendor data (black, dotted) with rule of 30 bounds (red, green).

As the verification performance is preserved, the proposed schemes are further examined regarding requirements of the biometric information protection standard [44]. This examination is performed in terms of [283], stating the following criteria: *i) only the client device can have access to the plain probe features, ii) the plain reference should not be seen by the client device, and only its encryption should be stored or handled during verification, and iii) the score should also be protected in order to prevent hill-climbing and inverse-biometrics attacks.* Firstly, both employed homomorphic Paillier cryptosystems provide semantic security: only secret keys are able to derive the plain probe after encryption, whereas the client device solely communicates encrypted auxiliary matrices ($\text{enc}_{pk1}(c_1)$) and $\text{enc}_{pk1}(c_2 + c_3)$, or the encrypted score ($\text{enc}_{pk}(S_{2Cov}(X, Y))$). Secondly, biometric references are communicated from the controller database server to the client device in the encrypted domain, assuming that the authentication server is able to protect the secret key $sk1$ and that no other entities will be able to put the protected biometric information into relation (to its plain-

Table 7.4: Contribution: complexity analysis for the proposed PLDA/2Cov HE schemes (verification) with the data sizes of the exemplary employed system ($p = 64$ bits, $v = 0.5$ KiB, $D = 250$).

Comparator Protection	2Cov subject	2Cov subject & vendor
N° encryptions	1	D^2
N° decryptions	1	$2 D^2 + 1$
N° additions	$4 D (D - 1)$	0
N° products	$4 D^2 + 2 D + 1$	$5 D^2 - 1$
N° exponentiations	$2 D$	$4 D^2$
Plain reference size	$p D$ ≈ 2.0 KiB	$p D$ ≈ 2.0 KiB
Protected reference size	$v (D + 1)$ $= 125.5$ KiB	$v (D^2 + D)$ ≈ 30.6 MiB
Plain comparator model size	$2 p D^2$ ≈ 1.0 MiB	$2 p D^2$ ≈ 1.0 MiB
Protected comparator model size	— —	$2 v D^2$ ≈ 61.0 MiB
Channels: amount of protected data exchanged	$v (D + 2)$ $= 126.0$ KiB	$v (5 D^2 + D + 1)$ ≈ 152.7 MiB

text by using this secret key, which is inaccessible to them). Similarly, the vendor data is protected in the sense that the vendor authentication server is assumed to be able to protect sk_2 . Finally, scores are computed in the protected domain and can solely be decrypted using secret key sk_2 . The irreversibility criterion is consequently met. Renewability is granted as depicted in [283]: if references are lost, new key pairs can be generated for the purpose of re-encrypting the database, such that i) re-acquisitions of enrollment samples are avoided when revoking corrupted references, and ii) comparisons of corrupt to renewed references result in non-matches, granting security and data privacy. Thus, references can easily be revoked, providing data privacy. Unlinkability is granted due to the probabilistic nature of the Paillier cryptosystem, where random numbers are used for different encryptions, i.e., in order to encrypt the same data Y twice, two different random numbers s_1, s_2 are drawn, such that: $enc_e(Y, s_1) \neq enc_e(Y, s_2)$, cf. [283, 287].

7.2.7 Complexity Analysis

In terms of complexity, each approach can be analyzed regarding the amount of required resources, i.e., the number of operations performed in the encrypted domain as well as the size of encrypted data sent over a channel. For a single verification attempt, the ciphertext

channel bandwidth is $\nu = 2n$ due to the Paillier ciphertext length in modulo n^2 domain [283], i.e., $\nu = 4096 \text{ bits} \frac{1 \text{ KiB}}{8 \cdot 1024 \text{ bits}} = 0.5 \text{ KiB}$ for the examined system. Tab. 7.4 summarizes the complexity of the proposed HE schemes. Regarding an i-vector dimension $D = 250$, the cosine HE approach requires $\nu D = 125 \text{ KiB}$ for storing a reference i-vector. For transmitting the protected score to the authentication server, 0.5 KiB are necessary, i.e., a protected scalar. The subject protective 2Cov HE scheme stores a reference tuple with $\nu(D + 1) = 125.5 \text{ KiB}$, communicating a protected scalar to the authentication server as well. However, the subject and vendor protective 2Cov HE scheme stores protected auxiliary matrices, requiring $\nu(D^2 + D) \approx 30.6 \text{ MiB}$. Therefore, the channel between client device and authentication server considers two protected matrices, requiring $2\nu D^2 \approx 61.0 \text{ MiB}$. The same holds for the channel between the vendor database and the operator authentication server. Regarding the protected data exchanged via the communication channels, the first proposed scheme comprises $\nu(D + 2) = 126 \text{ KiB}$ as the protected reference and score are transmitted. The second proposed scheme demands higher requirements: as the model hyper-parameters are protected, the client device to authentication server channel transmits auxiliary matrices comprising $2\nu(D^2) \approx 61.0 \text{ MiB}$. The same amount of data is loaded for the protected model from the vendor database server. Finally, a protected score is transmitted to the vendor authentication server, making application decisions. Conventional security protocols can be employed afterwards.

By employing HE, *data privacy* is protected: non-HE comparisons solely carried out by biometric service operators permit operators to utilize plaintext biometric features for other (non-biometric) purposes. HE prohibits operators to exploit *data privacy* by preserving *unlinkability*, *renewability*, and *irreversibility*. One may additionally assure *data security* by utilizing transport layer security (TLS), e.g., with RSA.¹⁴²

7.3 SUMMARY

In this chapter, PAD security towards unit-selection attacks and architectures ensuring privacy and data protection were examined. Countermeasures for unit-selection attacks were examined in a cross-language setup, detecting natural versus non-natural speech transitions based on (unfiltered) frequency features based on wavelet (DWT) and Fourier analyses (FFT). These unfiltered frequency-domain features were found to be feasible for PAD even when data and language

¹⁴² This method preserves privacy and protects data, it does not aim at security for machine communication. Exemplarily, an attacker could use the public key to encrypt a very high similarity score and send this score to an authentication server to get a positive verification outcome. The maintenance of data security (in machine communication) while preserving privacy (of biometric information) and protecting data (sensitive to vendors) is left to future work.

shifts occur between development and evaluation data. Effective unit-selection voice PAD countermeasures were proposed by examining DWT and FFT properties of probe samples in a single-system fashion. Therefore, neither speech production nor speech perception theory was necessary. SVM and GMM classifiers were found to be capable of distinguishing bona fide and unit-selection attacks samples. Further research on the frequency analysis of unfiltered speech signals seems promising.

Regarding privacy and data protection in state-of-the-art speaker recognition, homomorphic information protection was made available, especially to generative comparators (comparators employing statistical models). In contrast to the proposed approach, related work in biometrics solely considers non-generative comparators, such as XOR, DTW, Euclidean distance, and cosine similarity. Extending the HE scheme for cosine similarity comparison, biometric information protection is made available to the 2Cov comparator in two architectures. These architectures are used for comparators reporting on the evidence of probabilistic and discriminative embeddings (i-vectors and x-vectors). The first proposed HE architecture solely emphasizes on the protection of embeddings, which can be sustained under a fair complexity trade-off. By contrast, the second proposed HE 2Cov scheme provides subject and vendor data protection. However, the required complexity increases by a quadratic term. By pre-loading both protected comparator model parameters, the channel bottleneck is reduced to $\sqrt{3F^2 + F + 1} \approx 91.7 \text{ MiB}$ for a single verification attempt. This, however, limits the application scope to well-equipped infrastructures, e.g., call center and forensic scenarios. Depending on the application scenario, protected references may also be pre-loaded, further reducing the overall transmitted data amount to $\sqrt{2F^2 + 1} \approx 61.0 \text{ MiB}$. For mobile device voice biometrics, one may prefer to employ the first proposed architecture. Both approaches ensure biometric information protection requirements as of the ISO/IEC 24745 standard.

As the proposed schemes target 2Cov as prototype generative comparators, i.e., the full-subspace Gaussian PLDA special case, extensions to other members of the PLDA family and related comparators can be easily developed, especially as the bilinear PLDA computation comprises six matrix dot products (beside sums). Accounting for i-vectors not only as single point estimate features but also as latent variables, uncertainties associated to the single point estimation can be incorporated as well, e.g., targeting full-posterior PLDA. Also, HE schemes seem promising for end-to-end neural network system architectures, as the inner Frobenius product is computable in the protected domain. Extensions of the proposed architectures and implementations of alternative HE schemes are left to be the subjects of future work.

CONCLUSION

This dissertation addressed the problem of *speaker recognition in unconstrained environments* using the [Bayesian decision framework \(BDF\)](#), primarily in decision making for voice biometric systems with applications in smart home, online banking and payment, and forensics. A taxonomy for the visual performance evaluation considering the BDF has been proposed, interrelating latest advances of the forensics and speaker recognition communities in order to bridge gaps towards the biometrics community, particularly its standardization, but also to other machine learning disciplines targeting binary decision problems. The formalized way of decision making introduced by the BDF builds on the computation of [log-likelihood ratio \(LLR\)](#) scores and (if not possible, on) the calibration of system outputs to LLRs. In *unconstrained environments*, effects such as speech duration and background noise cause performance degradations. These impacts are examined regarding their *biometric distinctiveness* in terms of the security level speaker recognition systems are capable of supporting. By quality-based score normalization, re-calibrating degraded system outputs to LLRs, scores become optimizable for decision making without knowing decision requirements. Sustaining the LLR score property, decision making is formalized; magnitudes of belief ratios in decision priors and costs are definable, resulting in LLR thresholds that are comparable to well-calibrated scores. For the purpose of denoting and assessing decision policies, the concept of *verbal tags* is transferred from the forensic science community, when unable to compute LLRs, to performance evaluation, when unable to define decision policies. A guideline has been proposed in this dissertation for practitioners on denoting and refining LLR thresholds. Decision trade-offs are interrelated with the visualization of error rate trade-offs. Furthermore, as voice biometric systems are subject to attacks at the biometric sensor, but also in stored biometric data and at the comparator, countermeasures against replay attacks are proposed. Privacy preserving as well as data protective architectures are proposed based on [homomorphic encryption \(HE\)](#). Gaps are not only bridged between biometric information protection and speaker recognition; preservation of the LLR score property is made available to (biometric) cryptosystems. Neither client devices nor data controller nor [biometric system operators](#) nor [vendors](#) are able to relate references and probes. Each entity operates on protected data, i.e., operators compute encrypted scores without being able to interpret the system parameters distributed by vendors and client devices process encrypted biometric data.

8.1 MAIN RESULTS AND CONTRIBUTIONS

The main results and contributions of this dissertation are:

To the Theoretical Framework

- **Binary decision error trade-off (BET)** plots are proposed. Distances between two coordinates of the BET canvas resemble changes in the magnitude of prior and cost belief ratios (in LLR thresholds). The BET plot reveals betting log-odds of decision policy parameterizations. Decision policies are modeled, not error rates.

Conclusion: BET plots are applicable to the evaluation of *any* machine learning problem with binary decision outcome, as the latent decision subspace is revealed, considering LLRs to encode all comparison information in their value.

- The need to visualize *verbal scales of least-favorable decisions* in error trade-off plots is identified for the purpose of making LLRs digestible in performance reporting.

Conclusion: The outline of verbal scales for encompassed LLR values depends on the use-case, e.g., smart home applications might demand narrower but many LLR scales, whereas forensic case work might require wider but fewer scales.

- A requirement communication scheme is defined regarding BDF thresholds is to be met by the strength of evidence reported in the form of LLRs.

Conclusion: Formalized decision making is made available to requirement specifications. LLR values $\in [-5, +5]$ might already satisfy the vast majority of application requirements, as LLR values of 4.6 and 6.9 represent *one over a hundred* and *one over a thousand* (trivialized). Seeking higher LLR values increases the demand for evaluation data and discrimination performance.

- On any 2D error rate canvas, such as **receiver operating characteristic (ROC)** plots, thresholds *by error rate trade-offs* change when interpolating between them. Thresholds *by trade-offs in ratio magnitudes of prior and cost beliefs* (by beliefs) change when their canvas representation changes. That is, for well-calibrated system score outputs, changing between line segments of the **ROC's convex hull (ROCCH)**. For decisions by error rate requirements, score thresholds interpolate between empirical error rate coordinates. For decisions by belief requirements, LLR thresholds interpolate at angles of ROCCH edges.

Conclusion: Thresholds motivated from score observations (by error rates) are *fundamentally* different from thresholds formally derived from decision requirements (by beliefs).

- To **empirical cross-entropy (ECE)** plots, **normalized ECE (NECE)** plots are defined as an extension. This definition is analogous to the normalization of **applied probability of error (APE)** plots.

Conclusion: **NECEs** plots visualize information performance comparatively to a coin toss; C_{llr} and C_{llr}^{\min} values are read-off.

- The proposed taxonomy of performance evaluation differentiates between criteria (errors, information), types (discrimination, calibration), and audiences (analytic, reporting).

Conclusion: Gaps between the speaker recognition, biometrics, forensics, and machine learning communities are bridged; outlining implications of the **LLR** paradigm to evaluation.

To Novel Measures

- A measure is identified on **voice activity detection (VAD)** decision robustness and coherence in the speaker recognition task, facing noisy scenarios.

Conclusion: Pre-comparison figures of merit on (vision/audio) data segmentation decision robustness might aid technology evaluations to first harmonize data segmentation across participants, then to benchmark systems.

- The information accumulation of biometric voice references is demonstrated by increasing duration, sample completeness, in terms of the relative information conveyed in the acoustic **intermediate-sized vector (i-vector)** feature domain.

Conclusion: For i-vector point estimates, gaps are bridged to the biometric distinctiveness of other modalities. In unconstrained environments, features are extracted with uncertainty about their (expected) value, forming feature distributions form for **subjects**. Distinctiveness is more precise when the extraction uncertainty is propagated. Technology evaluations might benefit from employing pre-comparison figures of merit on the information provided by feature extraction, regardless of the recognition task (such as voice or face recognition) *in the wild*.

- **Quality vector (q-vector)** are proposed, motivated by **unified audio characterization (UAC)** and **quality measure functions (QMFs)**. Duration and noise conditions are exemplarily examined and quality conditions are linearly sampled. On acoustic i-vector features, q-vector conditions are found to be independent. Q-vectors are applicable to condition classification and to link related conditions from the perspective of signal processing by posterior probabilities, assuming flat priors.

Conclusion: Practitioners might employ non-flat priors to model the belief in a specific operative environment. The advantage of linking quality conditions by q-vectors rather than first classifying conditions and then employing condition-depending processing is the avoidance of condition *binning* errors. When discontinuing system architectures to follow the BDF, [pool adjacent violators algorithm \(PAV\)](#) based score calibration is limited to each condition bin. The empirical range of ideal LLRs is limited towards the amount of comparisons within each condition. However, if not subdividing quality by bins, scores over all conditions are used for training score calibration, and wider LLR ranges might occur as a result. Quality binning might result in a narrower range of empirically supported decision trade-offs (by magnitudes).

To Novel Methods

- The pre-selecting of cohort subsets is proposed based on quality by the divergence of q-vectors. By reflecting quality from the acoustic feature space perspective, pre-selection schemes that are based on (*oracle*) knowledge about present quality conditions are outperformed. This holds true especially for challenging conditions, as errors arising from condition binning are avoided.

Conclusion: Quality informed score normalization might aid other (biometric) recognition tasks.

- Score calibration based on [function of quality estimate \(FQE\)](#) is proposed, employing q-vectors. Comparison scores are re-biased depending on the (cosine) distance between q-vectors of the reference and probe. The yielded performance is competitive to the related work.

Conclusion: For samples of high quality, conventional calibration schemes yield better discrimination and calibration performance. Practitioners might distinguish between quality conditions of negligible uncertainty due to high precision and medium/low quality with uncertainty propagation.

- Based on the knowledge about noise to training score normalization with deep learning (mapping scores, quality information, and features to more discriminative scores), the availability of a wider-range of [signal-to-noise ratio \(SNR\)](#) levels (despite unavailability of biometric noise during calibration training) is identified to yield more robust recognition performance than the availability of non-biometric and biometric noise types (at unavailable low-SNR levels during calibration training).

Conclusion: The relevance of noise types might be pre-screened before seeking to cope a larger noise type variety, e.g., by further targeting reverberation and Lombard effects. Such robustness analyses might limit the number of conditions necessary to investigate when targeting *unconstrained environments*.

- **presentation attack detection (PAD)** is demonstrated between natural and non-natural speech of unit-selection attacks. Unnatural transitions between sound units are detected by classifying wavelet and Fourier components with support vector machines and **Gaussian mixture models (GMMs)**.

Conclusion: Effectively, research on **PAD** is a hill-climbing from either side (attacks and countermeasures). Ongoing research is necessary. For the field of speaker recognition, other research fields in speech communication are suitable to simulate **presentation attack instruments (PAIs)** created by experts.

- **Homomorphic encryption (HE)** schemes are proposed for **two covariance model (2Cov)** and **probabilistic linear discriminant analysis (PLDA)** comparators, preserving data privacy for **subjects** and ensuring data protection for **biometric system vendors**. LLRs are computed in the protected domain. Privacy and data protection are provided without losses in discrimination or calibration performance.

Conclusion: The proposed protocols assure the usability of LLRs in compliance to international standards and to the 2016 *European General Data Privacy Regulation*. By extending the protocol to also protect the data of vendors, novel licensing schemes could evolve. As encrypted data is licensed, the private key held by a vendor is revocable at any time. Before model parameters are distributed in plaintext to biometric service providers, and data is protected by expiring licenses, the illegitimate use after expiration presents a risk. By implementing a public key infrastructure between biometric service providers and comparison subsystem vendors, vendors hold the private key that is necessary for the operative status of a provider's application. Vendors are in full control of the operability of their distributed software but need to guarantee to preserve the availability of a provider's service. Therefore, vendors could also simply delete their private key when licensings discontinue. From the proposed 2Cov cryptosystem, solutions for other members of the PLDA family are directly derivable. They are also applicable to other classification tasks in machine learning. Unprotected systems can currently be upgraded with the proposed cryptosystem anytime without additional efforts in recapturing biometric enrolment data.

8.2 FUTURE PERSPECTIVES

A number of research lines arise from the work carried out in this dissertation. The following groups of specifically interesting considerations are identified as perspectives to the theoretical framework, novel measures, and novel methods.

To the Theoretical Framework. The understanding of the non-linearity in PAV-LLR score calibration is promising to investigate. The formal exploitation of the latent decision subspace might provide models on how to design experimental setups in machine learning, i.e., the amount of class \mathcal{A}, \mathcal{B} predictions to be conducted in technology evaluation. For one, as PAV-LLR segments empirical score sets into groups of alike LLR representations, group homomorphisms could be exploited in order to model the uncertainty of each LLR group given an experimental setup and to define an experimental setup that provides high precision for certain LLR ranges. For another, the binary-decision LLR is a special case of the multi-class LLR, also constituted by C_{llr} being a special case of the multi-class [strictly proper scoring rule](#) (see language recognition). Investigations on proper scoring rules for multi-class LLRs seem promising to better understand (biometric) identification recognition tasks. Identification is either a closed-set or an open-set task, the formulation of such proper scoring rules might address uncertainty about identity class representations in the feature space. These proper scoring rules on *identification LLRs* would be useful, in principle, LLRs encode the relation of feature space representations in their value—the LLR of the LLR is the LLR [39]. The quote summarizes a major property of LLR scores for binary (two-class, not yet multi-class) decisions: LLRs encode the class distribution within the feature space in their value. At the same time, the class distribution within the score space is encoded by the LLR value. For identification applications, LLR values could also encode the relation of multi-class decisions by their value. Eventually, the outline of multi-class comparators might benefit from good decision making in identification scenarios.

To Novel Measures. In future theoretical work, figures of merit on the appropriateness of an experimental design could be outlined, either targeting specific applications or application-independent constraints. The technology transfer of [q-vectors](#) to other (biometric) recognition tasks appears promising, as QMFs are already applied in face recognition [321]. Considering privacy concerns, e.g., for telephone conversations, end-users might want to countermeasure biometric espionage by automatically altering their speech signals in order to disguise or conceal their biometric identity. Conventionally, these countermeasures serve as [presentation attacks](#). In case of an illegitimate use of voice biometrics to exploit private conversations, however, they serve for the sake of preserving privacy in non-biomet-

ric scenarios. Such privacy preserving methods would need to sustain *speech intelligibility* as well: figures of merit depicting trade-offs between PAD performance and the *naturalness* of speech data might serve as future machine learning objectives.

To Novel Methods. In the *speaker recognition evaluation* (SRE) series of the *US National Institute of Standards and Technology* (NIST), the vast majority of data is provided as 8 kHz sampled English telephone speech data, transmitted over telephone communication channels, encoded as 8 bit μ -law files (with *a-law* encoded subsets in the 2016 and 2018 NIST SREs). As the usage of mobile devices has exaggeratedly increased within the past decade, the impact of higher bandwidths, video and speech codecs, and communication networking protocols deem to be interesting for research on mobile communication and video conference data. One could consider to employ such conditions in the q-vectors estimation. More types of intra-class variance appear promising to extend the scope of q-vectors to more operational scenarios, where voice quality is degraded by, e.g., reverberation [38], Lombard effects [322], and aging [323]. For training q-vector extractors, the additional use of the well-studied PRISM dataset [40] appears promising, especially as comparability to other related work is provided, when targeting a broader diversity of speech signal distortions (like reverberation). q-vectors could aid cross-language solutions by language-independent generalizations, such as systems that are trained on English speech data but employed on German and French speech data. Language and dialect adaptive calibration could be investigated as q-vectors might be capable of linking dialects by their representations within a system's signal processing. Multi-algorithmic system fusion on the score level might benefit from employing q-vectors, UACs, and QMFs [324, 325] in order to sustain high discrimination performance in *unconstrained environments*, while sustaining low calibration losses. It seems promising to not only investigate the quality-adaptive re-bias of thresholds/scores but also their re-scaling using comparison depending quality information as well as combinations of re-bias and re-scaling.

To further advance privacy preservation and data protection, the gap between secure two/multi-party computation and speech technology could be bridged more, e.g., by using the *Yao's garbled circuits protocol* [326]. The cryptographic community already provides frameworks for implementing secure two/multi-party computations using garbled circuits, HE, and combinations of both (among others), e.g., see the *TASTY* framework [327] and the *ABY* framework [328]. The former provides non-linear functionality (e.g., comparisons) based on which more flexible privacy-preserving PLDA/2Cov protocols are possible. The latter accommodates alternative solutions to HE, e.g., multiplications carried out via arithmetic sharing based on symmetric cryptography (whereas the Paillier HE uses asymmetric cryptog-

raphy) and on *oblivious transfer* [329, 330]. As an alternative to Paillier HE, whose security is broken when factoring large integers is efficient, post-quantum secure lattice based fully HE seems promising for preserving privacy in speaker recognition (e.g., see the *PALISADE* library [331]).

Regarding speech segmentation by VAD, research is relevant to sustain the robust estimation of q-vectors. On the one hand, VAD decision errors are propagated throughout biometric comparison and to the q-vector estimation. On the other, an uncertainty about the VAD decision could be formalized in a marginal estimation of q-vectors. Moreover, fast and robust VAD on real-world speech data is not only relevant to speaker recognition (and its relation to impact these systems [332]), but also to other fields in speech communication, such as *automatic speech recognition* (ASR). The outline of end-to-end deep neural network (DNN) architectures might employ layers inspired by q-vectors in order to produce robust LLRs, perhaps as a variation of the attention mechanism [154].

Finally, computations of LLRs with DNNs remain an interesting research topic, as DNNs are, in principle, capable of estimating any function mapping. Inference using DNNs is not yet fully solved: discriminative embeddings are not interpretable, whereas conventional models and approaches of signal processing are. Among the conventional machine learning fields like *variational Bayes*, two directions seem promising to follow: the extraction of *meta-embeddings* [54, 99] and the thoroughly principled treatment of probabilistic input data using *mixed sum-product networks* [333] and *automated Bayesian density analysis* [334]. The first idea is based on conventional DNNs, enforcing them to yield outputs, serving as meta-descriptors of reference and probe embeddings, capable of computing LLRs by a principled uncertainty propagation throughout a DNN. The latter two ideas enforce a probabilistic data treatment throughout DNN layers in order to estimate likelihoods for Bayesian inference, while being capable of reporting data anomalies. Such LLR estimating DNNs may be outlined for verification and identification recognition tasks, making them appealing to serve in commercial and forensic scenarios as the principled propagation of uncertainty and probabilities preserves interpretability of decisions based on DNN outputs.

A Final Remark. In performance assessment by error rates, the achievement of limiting error occurrences prevails. By contrast, in performance assessment by information, the formal definition of decision requirement beliefs (e.g., class occurrences and cost trade-offs) is accommodated with decision risk assessment on the benefit of using the evidence reported by systems for one single or multiple settings. Eventually, the elaboration on implications of LLRs from Frequentist and Bayesian perspectives might serve well in further harmonization of theory and technology across communities.

GREEDY PLDA-RBM AND MOBILE ENVIRONMENTS

This section investigates¹⁴³ the discrimination power of [probabilistic linear discriminant analysis \(PLDA\)](#), which (as a graphical) is reformulated to the graphical representation of a restricted Boltzmann machine (RBM). By using RBMs, deep neural network (DNN) architectures can be formulated for PLDA. This section focuses on limited availability of DNN training data, particularly, to limited mobile speech data, see chapter 3. In commercials, small medium enterprises (SMEs) can be confronted with limited training data. In forensics, speech signal properties are often distinct for each case (if resources for appropriate data collection are available, these collections are limited by the availability of the resources). As such, the following research question is raised:

Not all [vendors](#) of comparison subsystems are able to develop systems on datasets providing thousands of hours of speech. Can PLDA classifiers thus be de-noised on limited mobile device speech training data?

In [336], a proof-of-concept study proposes PLDA-RBM as an approach, where units representing biometric (speaker) and non-biometric (channel) factors are factorized, achieving a comparable performance to PLDA. For this investigation (carried out on the MOBIO dataset), mobile environments with limited training data are targeted. Emphasis is put on the suppression of channel effects and the recovery of subject discriminative information for the training of comparators on a small dataset. The motivation of this study is drawn from the motivation of the [intermediate-sized vector \(i-vector\)](#) paradigm that originated from joint factor analysis (JFA) [127, 130]: discriminative information is still observed in JFA decomposed non-biometric factors [130]. In this section, the following hypotheses are investigated:

- As limited training data is available, the conventional Gaussian assumption for hidden layers might be insufficient, whereas a Bernoulli assumption might better account for the binary decision task.
- Biometric information can be purified by exploiting (supposedly non-biometric) channel units rather than (supposedly biometric) speaker units when greedy architectures are employed.

¹⁴³ Parts of this section are based on a collaborative work with Hong Hao, Themis Stafylakis, Christian Rathgeb, and Christoph Busch [75], which emerged from the collaboration with Hong on his master's thesis [335].

A.1 PLDA WITH RESTRICTED BOLTZMANN MACHINES (RBMS)

RBMs are used to distinguish between biometric and non-biometric factors of PLDA in a JFA-like fashion.

A.1.1 *Restricted Boltzmann Machines*

An RBM is a bipartite undirected graphical model with no connections between units of the same layer [337]. This property makes the distributions of the two layers conditionally independent and therefore allows the application of fast sampling-based training techniques. RBMs can serve different purposes, e.g., probabilistic principal component analysis (PPCA), feature reconstruction, and unsupervised initialization of DNNs [336, 338–340].

RBMs are two-layer structure models containing a visible and a hidden layer $\mathbf{v} = \{v_i\}_{i=1,\dots,d_v}$, $\mathbf{h} = \{h_j\}_{j=1,\dots,d_h}$ with the numbers of visible and hidden units d_v , d_h , which are connected through a weight matrix $\mathbf{W} = \{w_{i,j}\}$ [336, 338]. The joint probability density function (pdf) of (\mathbf{v}, \mathbf{h}) is a $(d_v + d_h)$ -dimensional Gaussian that depends on an energy function $E(\mathbf{v}, \mathbf{h})$:

$$P(\mathbf{v}, \mathbf{h} | \mathbf{W}) = Z^{-1} e^{-E(\mathbf{v}, \mathbf{h})}, \quad (\text{A.1})$$

where Z is the normalizing constant to sustain that the area under this pdf is equal to one.

Energy functions model the distribution of visible and hidden units: *Gaussian-Gaussian* (GG) layers assume Gaussian distribution for visible and hidden units, *Gaussian-Bernoulli* (GB) layers assume Gaussian distribution for visible units and Bernoulli distribution for hidden units. By assuming zero mean for GG energy functions, the distributions take the form of PPCA [336, 339]. The GB energy function $E(\mathbf{v}, \mathbf{h})$ considers the hidden unit pdf to be Bernoulli-distributed [339, 341], such that the GB energy function takes the form of [339]:

$$E(\mathbf{v}, \mathbf{h}) = \sum_{i=1}^{d_v} \frac{v_i^2}{2\sigma_i^2} - \sum_{i=1}^{d_v} \sum_{j=1}^{d_h} b_j h_j - \sum_{\substack{i \in d_v \\ j \in d_h}} \frac{v_i}{\sigma_i} h_j w_{ij}. \quad (\text{A.2})$$

A.1.2 *PLDA-RBM Algorithm*

RBMs are used in a specific way to graphically represent the PLDA comparator. Visible units representing i-vectors are decomposed into hidden speaker units h_s^{speaker} and hidden channel units h_c^{channel} , respectively representing (biometric) speaker and (non-biometric) channel/residual factors. Fig. A.1c depicts the basic idea: during enrolment, speaker-dependent RBM weights $\mathbf{W}(s)$ are learned with the

constraint to purify h_s^{speaker} . During verification, the same weights $W(s)$ are used for purifying probe i-vectors [336].

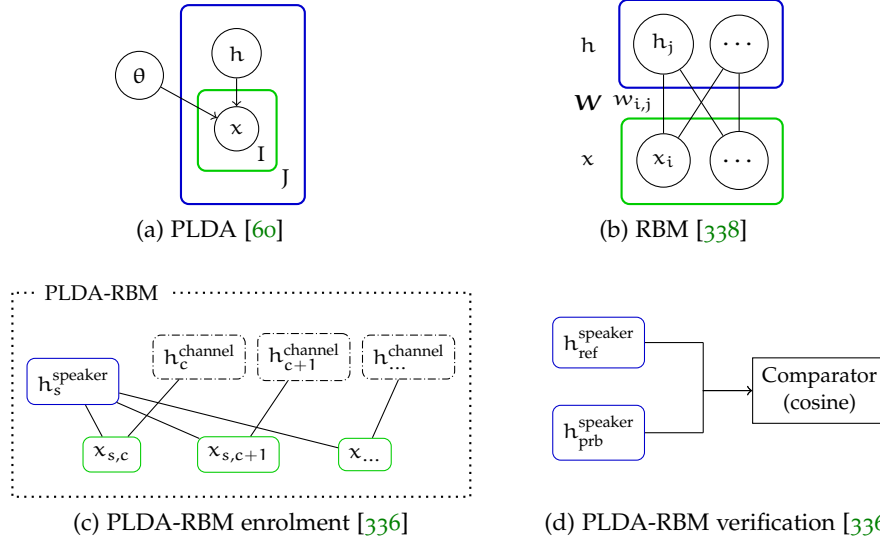


Figure A.1: Factorization concepts with data samples x , hidden variables h , PLDA parameters $\theta = \{\mu, \Phi, \Lambda\}$, and weights W .

The main difference between the presented approach and [336] is the usage of Bernoulli hidden layers, i.e., a GB PLDA-RBM. The PLDA-RBM is trained with *single-step contrastive divergence* (CD1), using mini-batches and standard L2 regularization, while no momentum terms are added [339]. During recognition phase, features are extracted using the speaker layer (one per i-vector). In the case of multi-sample enrolments, references are created by taking the average representation. Reference and probe features are compared by the cosine distance. Fig. A.1 depicts PLDA, RBM, and PLDA-RBM architectures.

A.2 CONTRIBUTION: DEEP PLDA-RBM DESIGNS

One motivation behind the i-vector paradigm is the insufficiency of JFA in distinguishing between (biometric) speaker and (non-biometric) channel information, as channel factors were shown to containing speaker information [130]. Incorporating both, a two-step approach emerged: an acoustic total variability subspace is estimated first, then a biometric subspace is estimated, where biometric comparisons are carried out (see section 2.5). In cases of limited labeled training data, however, the problem of speaker information linkage to channel factors may reappear.

A.2.1 Greedy Architectures

To address this issue, a deep architecture is proposed in which the channel factors of the initial PLDA-RBM are further processed using a second PLDA-RBM model (greedy training). The same approach is repeated N times, leading to a deep architecture that is trained using greedy CD1. For completeness, the same idea is also applied to the speaker layers. Both approaches are described below.

A.2.1.1 Stacking on Channel Units

Following the hypothesis of *biometric information to still be present in hidden channel units*, the extracted hidden channel units are further examined by deeper PLDA-RBM layers, i.e., hidden channel units of the $(N-1)^{\text{th}}$ -layer are re-processed by the N^{th} -layer, resulting in the hidden speaker units $\hat{h}_s^{\text{speaker}}$. Thereby, CD1 training is performed layer-wise (greedy). L2 regularization is only applied on the first layer since the weights of deeper layers decrease dramatically in this experimental setup on limited training data. For stacking on channel units, a feature fusion of the hidden speaker units of all layers $\{h_s^{\text{speaker}}, \dots, \hat{h}_s^{\text{speaker}}\}$ is proposed by concatenation in order to assemble an augmented reconstructed biometric feature, cf. Fig. A.2a.

A.2.1.2 Stacking on Speaker Units

Following the hypothesis of *noisy speaker units*, the reconstructed hidden speaker units are refined by deeper PLDA-RBM layers, i.e., hidden speaker units of the $(N-1)^{\text{th}}$ -layer are re-processed by the N^{th} -layer, resulting in the hidden speaker units $\check{h}_s^{\text{speaker}}$, which is proposed as biometric features, cf. Fig. A.2b. However, this approach might also lead to a further loss of biometric information if the original h_s^{speaker} units already comprise well-reconstructed features that can be over-fitted by re-assessment, e.g., due to limited training data.

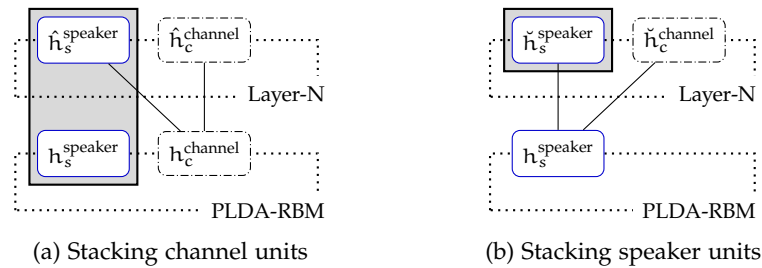


Figure A.2: Comparison of proposed deep PLDA-RBM designs with \hat{h} as deep hidden units of deeper layers N . Gray layers indicate the proposed biometric features.

A.2.2 Experimental Validation

The baseline PLDA-RBM system is based on the *Matlab Environment for Deep Architecture Learning (MEDAL)* [342]. PLDA-RBM layers are CD1-trained by using the background set, where the mini-batches comprise a quarter of the i-vectors per subject. Then, PLDA-RBM is re-trained using the development set in order to cope with dataset shifts on limited short-utterance mobile data.

Table A.1: Performance of baseline systems on development set.

System	Female			Male		
	EER	FMR ₁₀₀	C_{llr}^{\min}	EER	FMR ₁₀₀	C_{llr}^{\min}
G-PLDA [132]	15.3	63.5	0.488	12.2	44.6	0.413
PLDA-RBM						
GG	17.7	64.2	0.552	16.7	60.4	0.526
GB	13.5	51.2	0.451	12.3	48.3	0.418

Tab. A.1 indicates the baseline performance in terms of **equal error rate (EER)** (in %), **FNMR at a 1% FMR (FMR₁₀₀)** (in %) and C_{llr}^{\min} of G-PLDA with 400 speaker factors and PLDA-RBM with 400 hidden speaker units with GG and GB energy functions. Re-training is not applied at this stage. GB PLDA-RBM significantly outperforms GG PLDA-RBM. Performance gains to the G-PLDA baseline can also be observed. However, this observation may change on big data background sets, such as the **speaker recognition evaluation (SRE)** scenarios of the **US National Institute of Standards and Technology (NIST)**.

An optimal configuration regarding the number of hidden speaker and channel units is examined on the development set. Fig. A.3 depicts C_{llr}^{\min} for GB PLDA-RBM with development set re-training. Good results are observed at 850 hidden units.

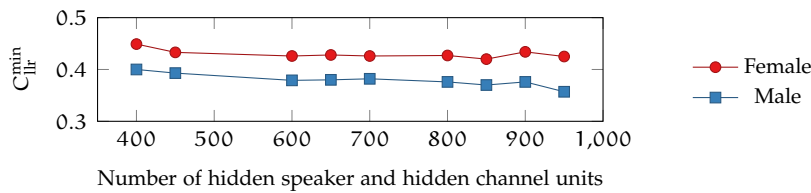


Figure A.3: Comparison of different numbers of hidden speaker and channel factors incorporating development set re-training.

The hypotheses are examined on up to three layers on the development set, cf. Tab. A.2. While stacking on hidden speaker units decreases information in terms of C_{llr}^{\min} , stacking on channel units retrieves information.

Tab. A.3 depicts C_{llr}^{\min} of channel unit stacked PLDA-RBM. In particular, the performance of hidden speaker units per layer is compared

Table A.2: C_{llr}^{\min} comparison of stacking concepts for hidden speaker unit extraction on up to three layers on the development set: channel units (channel-stacked) and speaker units (speaker-stacked).

# Layers	Female		Male	
	Channel-stacked	Speaker-stacked	Channel-stacked	Speaker-stacked
1		0.420		0.370
2	0.392	0.481	0.341	0.452
3	0.394	0.487	0.346	0.475

to the performance of the proposed concatenation of hidden speaker units (assembled from all layers). Subject information can still be retrieved on the fifth layer but without further significant gains. In this setup, hidden speaker units are more prone to be zero on deeper layers (to convey less biometric information).

Table A.3: C_{llr}^{\min} comparison of recovered i-vectors by the channel-stacked PLDA-RBM architecture on N^{th} -layer (layer) and layer-concatenated features.

		# Layers	1	2	3	4	5
Female	Layer			0.505	0.551	0.691	0.715
	Concatenated		0.420	0.392	0.394	0.398	0.394
Male	Layer			0.459	0.510	0.639	0.681
	Concatenated		0.370	0.341	0.346	0.340	0.342

Tab. A.4 compares the half-total error rate (HTER)¹⁴⁴ and C_{llr}^{\min} performances of the examined GB PLDA-RBM with 850 hidden speaker and channel units to female and male systems of the 2013 MOBIO SRE, particularly, to those systems with one acoustic feature extractor and without multi-algorithmic comparator fusion (to comparable systems). Contrary to state-of-the-art systems, either system compares [Gaussian mixture models \(GMMs\)](#) with [universal background models \(UBMs\)](#), following the traditional GMM – UBM approach. On

¹⁴⁴ HTER is the primary evaluation metric of the 2013 MOBIO SRE [37]. Computationally, the HTER equals the [decision cost function \(DCF\)](#) parameterized with $\tilde{\pi} = 0.5$ weights (with corresponding threshold $\eta_{\tilde{\pi}=0.5} = 0$), i.e., the maximum entropy of cost and prior ratios (e.g., unknown decision policy specifications). Contrary to the DCF, however, the HTER threshold is determined on a development set as the EER threshold. The EER threshold, in turn, is derived by the maximum value of all [minimum DCF \(minDCF\)](#) parameterizations (see section 2.4.3). In other words, the EER threshold can be different to $\eta_{\tilde{\pi}=0.5}$; a different threshold also corresponds to a different parameterization of the DCF criterion. Therefore, the HTER metric appears incoherent for poorly calibrated systems, since the EER threshold equals $\eta_{\tilde{\pi}=0.5}$ only for well-calibrated systems. In other words, where the DCF criterion penalizes (discrimination and) calibration performance on the evaluation set in a coherent manner, the HTER criterion penalizes calibration mismatches on the evaluation as well as on the development set (solely in a coherent manner if systems are well-calibrated).

Table A.4: HTER (in %) and C_{llr}^{\min} comparison of best single systems to the proposed systems on the evaluation set of the 2013 MOBIO SRE. (None of the systems incorporates calibration.)

System	Female		Male	
	HTER	C_{llr}^{\min}	HTER	C_{llr}^{\min}
MOBIO-female [37]	11.6	n/a	9.1	n/a
MOBIO-male [37]	12.8	n/a	8.9	n/a
G-PLDA (as of [132])	16.4	0.522	9.9	0.326
GB PLDA-RBM	12.0	0.397	10.6	0.361
2-layer GB PLDA-RBM (channel-stacked)	11.3	0.368	9.0	0.319

the evaluation set, these traditional systems outperform the conventional PLDA (with Gaussian assumptions), which is baseline to the proposed GB PLDA-RBMs. The two-layer (channel-stacked) GB PLDA-RBM architecture achieves similar performance as the GMM-UBM systems.

A.3 SUMMARY

PLDA-RBM benefits from GB assumptions on limited mobile data, outperforming the conventional G-PLDA by reconstructing speaker features and removing channel impacts. Moreover, deep PLDA-RBM is shown to recover relevant biometric information from discarded (supposedly non-biometric) channel units by using the proposed stacking on channel units concept. Compared to (comparable) systems of the 2013 MOBIO SRE (which rely on the GMM-UBM approach), the proposed system achieves similar results. This is particularly desirable (e.g., for forensic scenarios) when processing efforts are of minor concern but reliable evidence is rather important. Especially on female speech data (during comparator training, less female speech data is available than male speech data in the 2013 MOBIO SRE), performance gains are observed.

PAV-LLR CALIBRATION: STEP-BY-STEP EXAMPLE

Isotonic regression monotonically maps (uncalibrated) score groups of equal posterior prediction to their depending [log-likelihood ratio \(LLR\)](#) values. As such, (uncalibrated) scores are grouped by their contribution to decision making. Fig. B.1 illustrates two examples of the [pool adjacent violators algorithm \(PAV\)](#). The first example depicts selected iteration steps on ten class \mathcal{B} and seven class \mathcal{A} scores¹⁴⁵, see Figs. B.1a to B.1i. The second example (see Fig. B.1j) depicts the linking of (uncalibrated, synthetic) scores to their corresponding posterior probabilities (after PAV). Initially, PAV aligns class labels 0, 1 according to their depending score values by sustaining monotonicity (on same posterior values, zeros are aligned before ones). In each iteration (assessing scores from left to right in a decision-by-decision manner), a posterior group of a current score is determined. Eventually, previous posterior groups are updated in order to preserve monotonicity.

For **class \mathcal{A} scores**, the corresponding group represents the 100% posterior probability, cf. the transitions to the second, fifth or final iteration. Class \mathcal{A} scores either start new posterior groups or are associated to a previous posterior group, if their associated posterior probability is 100% (when class \mathcal{B} scores are not present in that group). For **class \mathcal{B} scores**, the corresponding posterior group resembles an update of the previous posterior groups. By observing one more class \mathcal{B} decision, a current group's posterior is lowered. Thereby, different posterior groups can collapse into one single posterior group (if the update leads to an equal or smaller posterior probability to its preceding group).

Example: PAV-LLR Calibration of 17 Scores

PAV score calibration is an iterative regression. This example depicts certain iterations for the calibration of 17 scores. In the first two iterations, class \mathcal{B} scores solely contribute to the group of 0% posterior probability. Then (iteration five), a class \mathcal{A} score with a following class \mathcal{B} score is observed. To sustain monotonicity, both scores are mapped to the group of 50% posterior probability. At iteration six, two consecutive class \mathcal{A} scores (which were in the group that represented 100% posterior probability at iteration five) are followed by one class \mathcal{B} score. The group of these scores is alleviated

¹⁴⁵ For class \mathcal{B} scores, the set is $\{-5, -4, -2, 0, 1, 2.5, 3, 3.5, 4.0, 4.25\}$; for class \mathcal{A} scores, the set is $\{-3, -1.5, -1, 2, 3, 4.5, 5\}$. The sorted score classes of this example are: $(\mathcal{B}, \mathcal{B}, \mathcal{A}, \mathcal{B}, \mathcal{A}, \mathcal{A}, \mathcal{B}, \mathcal{B}, \mathcal{A}, \mathcal{B}, \mathcal{B}, \mathcal{A}, \mathcal{B}, \mathcal{B}, \mathcal{A}, \mathcal{A})$. Both classes are observed for the score value of 3; the class \mathcal{B} label is sorted in before the class \mathcal{A} label.

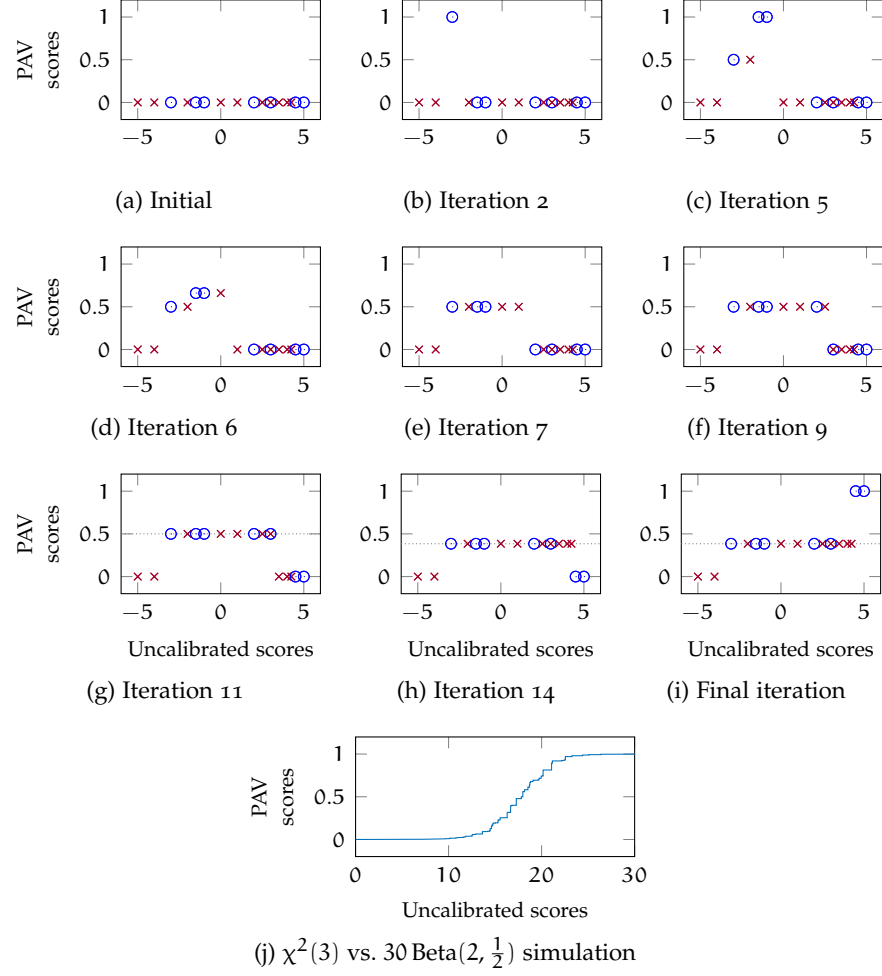


Figure B.1: Examples of the PAV-LLR algorithm with class \mathcal{A} scores (blue circles) and class \mathcal{B} scores (red crosses): (a) to (i) depict an example for 17 raw scores on distinct iterations, (j) depicts the final PAV mapping on 10^6 class \mathcal{B} and 10^4 class \mathcal{A} synthetic scores. The dotted line in the last row indicates the PAV score of the group with the most uncalibrated scores.

to the group of $\frac{2}{3}$ posterior probability. Then (iteration seven), another class \mathcal{B} score is observed, such that the previous group of $\frac{2}{3}$ posterior probability is further alleviated to the group of 50% posterior probability. As this group already exists as of iteration five, the union of both groups resembles (eight out of seventeen scores are processed, the posterior odds of the 50% group are 3 : 3). Until the final iteration of this example, more scores are augmented to this group of 50% posterior probability. First, the group is augmented by two scores; one of each class, another 50% (iteration 9). Second, two scores of different classes but same value are processed (iteration 11)—in this example, two scores of the value 3.0—, the group remains at 50% posterior probability (5 : 5 as odds). Then (iteration 14), three class \mathcal{B} scores follow, alleviating this score group further

(5 : 8 as odds, and $\frac{5}{13}$ as posterior probability). Finally, two more class \mathcal{A} scores follow, which resemble the new group of 100% posterior probability. LLRs are computed from PAV scores by Eq. (2.43), i.e., after removing the database prior $\hat{\pi}$ from the posterior scores. In this example, the database prior is $\hat{\pi} = \frac{7}{17}$, such that the LLRs of the 5 : 8 PAV group (representing most of these uncalibrated scores) have the value $-\log \frac{28}{25} \approx -0.11$. In the depicted example, most of the uncalibrated scores are assigned to the same group. These scores are calibrated to one LLR value. From the observation of uncalibrated score histograms, the latent decision space is inferred, in which (effectively) decision making proceeds.

Notably, the magnitude of these LLR groups depends on the empirical results. One might aim at systems that output LLRs on a host of scales, e.g., which provide *weak* to *extremely strong* support to either decision across all scales (see verbal scales and the scale of least-favorable decisions in section 4.2). The example above, however, illustrates that the sorting of class labels by empirical score outputs (class label permutations) limits the variety of LLR scales and therefore the extent of decision scales a system is capable of supporting. Exemplarily, the least-favorable *extremely strong* decision towards proposition \mathcal{A} requires an LLR of $\log \frac{10^6}{1}$. By reformulating Eq. (2.43), necessary values of PAV scores and database priors are exploitable. Exemplarily, when assuming a database prior of $\hat{\pi} = \frac{1}{1\,000\,001}$ (e.g., one class \mathcal{A} score and a million class \mathcal{B} scores), the PAV group representing 50% posterior probability would be calibrated exactly to the above-mentioned LLR. By contrast, when assuming a database prior of $\hat{\pi} = \frac{1}{2}$ (equal amount of class \mathcal{A} and class \mathcal{B} scores), the PAV group representing $\frac{1\,000\,000}{1\,000\,001}$ posterior probability (e.g., a million class \mathcal{A} scores with one higher class \mathcal{B} score) would be exactly calibrated to the above-mentioned LLR. Thus, depending on an experimental setup, (empirical outputs as) class label permutations appear to be more crucial: if, for a class label sequence (e.g., $\mathcal{A}, \mathcal{B}, \mathcal{A}, \mathcal{A}, \mathcal{B}$; odds 1 : 1 and 2 : 1, two posterior groups of $\frac{1}{2} = 50\%$ and $\frac{2}{3} \approx 67\%$) two class labels change, their scores might be differently sorted (e.g., due to insufficient sample quality). A permutation (e.g., $\mathcal{A}, \mathcal{A}, \mathcal{B}, \mathcal{A}, \mathcal{B}$; odds: 3 : 2, one posterior group of $\frac{3}{5} = 60\%$) differs in the posterior groups. The range of empirical LLRs depends on the split and merge of posterior groups. Nonetheless, for deterministic system outputs, the change of an evaluation setup is not changing the (non-empirical) LLR values. If a system is well-calibrated, empirical and non-empirical LLR values are identical, otherwise an empirical LLR value (from PAV score calibration) approximates the depending non-empirical LLR value (which is fundamentally linked to a comparator). For the latter, where the gaps between empirical and non-empirical LLR values are relevant, one could explore credibility intervals of PAV score calibration depending on experimental setups (see suggested future work in section 8.2).

Independent of the research and evaluation field, in which systems are assessed for binary decision making, finding credible intervals to the most likely PAV outcomes of an experimental setup appears very promising. By doing so, setups to yield the desired information drawn from research and evaluation studies could be outlined and the likelihood of remaining with the mere result “but more data is necessary” minimized.

The PAV-LLR algorithm is formally addressed in [28, 78, 79]. The family of binary *strictly proper scoring rules* and the multi-class logarithmic cost are addressed by *Niko Brümmer* in [116] and [28, appendix D]. By allowing PAV to result in posterior probabilities of 0% or 100%, infinite LLR values would be allowed. To prevent this, one could employ PAV with *Laplace’s rule of succession* (see [78, 79, 107] for details). Effectively, two dummy class labels are added front and back to the existing labels (\mathcal{A}, \mathcal{B} , then the actual class sequence, \mathcal{A}, \mathcal{B}). As such, unobserved scores are acknowledged to occur outside the range of observed scores (a minimum posterior larger zero and maximum posterior smaller one are sustained), while the posteriors assigned to groups of intersecting scores remain unchanged. As further documented in the code of [79], the (convex) *receiver operating characteristic (ROC)* in *detection error trade-off (DET)* and *binary decision error trade-off (BET)* plots are stopped from approaching the plot axes (on sparse data) and instead remain unchanged, i.e., in the well-populated regions. Fig. B.2 illustrates the impact of *Laplace’s rule of succession* on the visualization of *ROC’s convex hulls (ROCCHs)* in BET plots.

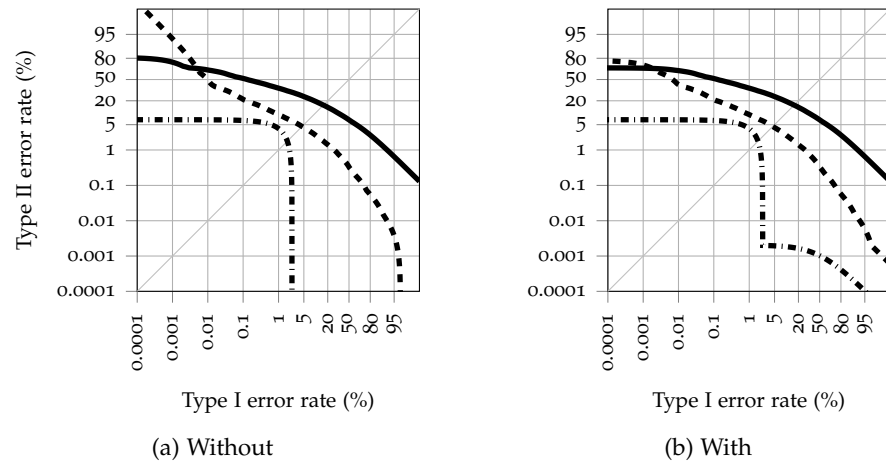


Figure B.2: Example of *Laplace’s rule of succession* impacting the visualization of ROCCHs (in BET plots), exemplarily computed from synthetic score sets (red versus green) sampled from 2.3a $\mathcal{N}(0, 1)$ versus $\mathcal{N}(3, 2)$ (solid), 2.3b $\chi^2(3)$ versus $25 \text{ Beta}(2, \frac{1}{2})$ (dashed), 2.3c $\mathcal{U}(-4, \frac{1}{10})$ versus $\text{Beta}(\frac{9}{10}, \frac{1}{2})$ (dash-dotted).

GLOSSARY

applied probability of error Bayes error rate for binary decision systems, cf. [28, 61], 59, 154, 247, see also: BDF, LLR & DCF

attack potential *measure of the capability to attack a target of evaluation given the attacker's knowledge, proficiency, resources, and motivation* [93]

attack presentation classification error rate *proportion of attack presentations using the same PAIS incorrectly classified as bona fide presentations in a specific scenario* [93].

Note: the APCER can be defined regarding a certain PAIS ($\text{APCER}_{\text{PAIS}}$) but also depending on an attack potential (AP) across multiple PAISs [93]:

$\text{APCER}_{\text{AP}} = \max_{\text{PAIS} \in A_{\text{AP}}} \text{APCER}_{\text{PAIS}}$, where A_{AP} is the set of evaluated PAISs, 35, 126, 222, see also: presentation attack, bona fide presentation, PAIS & attack potential

automatic speech recognition recognition using speech for non-biometric purposes of, e.g., detecting words, phrases, or semantically inferring the intention of natural speech, present in interactive voice response (IVR) systems, 81, 156, 172, 252

Bayesian decision framework a framework for decision making, combining Bayesian inference and decision theory, 4, 17, 24, 47, 59, 64, 70, 78, 108, 115, 127, 175, 192, 215, 245

binary decision error trade-off visualization of Type I versus Type II error rates in a quantile-quantile (Q-Q) plot that uses logistic distributions to model changes in formally denotable decision thresholds derived from parameterized priors and cost beliefs. The depending plot is motivated from the Bayesian decision framework, 14, 110, 128, 141, 223, 246, 266, see also: ROC, DET & BDF

biometric applicant *individual seeking to be enrolled in a biometric enrollment database* [1]

biometric attendant *agent of the biometric system operator who directly interacts with the biometric capture subject* [1], see also: biometric capture subject

biometric capture subject *individual who is the subject of a biometric capture process* [1], 11, 29, 165, 209, 215, 226, 235

biometric characteristics examiner *individual with authority to assess biometric characteristics and who does so for the purpose of resolving a biometric claim* [1]

biometric data controller *person or organisation which, alone or jointly with others, determines the purposes, means, and goals of the pro-*

cessing of biometric data [95].

Note: the term is motivated by the EU GDPR [272], 234

biometric data processor *person or organisation that processes biometric data on behalf of the biometric data controller* [95].

Note: the term is motivated by the EU GDPR [272], 234, see also: [biometric system operator](#), [biometric system vendor](#), [biometric system owner](#) & [biometric system provider](#)

biometric distinctiveness referring to the fact that biometric characteristics should be sufficiently different across subjects (comprising the population), cf. *biometric uniqueness* in [214], 163, see also: [subject](#)

biometric enrollee *biometric data subject whose biometric data is held in a biometric enrolment database* [1]

biometric operational personnel *individuals, other than the biometric capture subjects, who take an active role in the operation of the biometric system* [1]

biometric system end-user deprecated term of [1], 16, 27, see also: [biometric capture subject](#)

biometric system operator *person or organization that executes policies and procedures in the administration of a biometric system* [1], 4, 21, 27, 55, 78, 135, 140, 184, 211, 216, 234, 239, 245

biometric system owner *person or organization with overall accountability for the acquisition, implementation, and operation of a biometric system* [1], 27, 40, 135, 140, 184, 211, 216, 234, 239

biometric system provider *person or organization that supplies a biometric system to system users.*

Note: the term reflects a biometric system supplier as a service provider, 135, 140, 184, 211, 216, 234, 239

biometric system vendor *person or organization that creates and produces biometric systems or their components.*

Note: the term reflects a biometric system supplier as a creator of biometric system components, 4, 10, 16, 21, 27, 55, 71, 135, 140, 184, 211, 216, 227, 239, 245, 249, 255

bona fide presentation *interaction of the biometric capture subject and the biometric data capture subsystem in the fashion intended by the policy of the biometric system* [93], 215, 219, 221

bona fide presentation classification error rate *proportion of bona fide presentations incorrectly classified as presentation attacks in a specific scenario* [93], 35, 126, 222, see also: [presentation attack](#) & [bona fide presentation](#)

claimant *individual making a claim that can be verified biometrically* [1]

cooperative biometric capture subject *biometric capture subject motivated to achieve a successful completion of the biometric acquisition process* [1]

decision cost function weighted sum of error rates (weights not necessarily sum up to one) with weights depending on the parameterization of the Bayes risk; conventionally used in NIST SREs, e.g., with $\tilde{\pi} = 0.01$ in the 2010 NIST SRE [34].

Note: the 2012, 2016, and 2018 NIST SREs [22, 35, 343] employ more complex operating points, i.e., averages of two or three DCFs. The 2016 and 2018 NIST SREs [22, 343] also examine combined DCFs over 16 partitions, for which an average evaluation criterion is formed, 5, 59, 98, 118, 127, 177, 260, see also: BDF, NIST & SRE

detection error trade-off visualization of Type I versus Type II error rates in a quantile-quantile (Q-Q) plot that uses standard normal distributions, cf. [88], to ease visual comparability by modeling error trade-offs in Gaussian distributed scores as straight lines rather than curves. The depending Q-Q plot is motivated from the observation that cohort based score normalization tends to Gaussianize scores, 32, 33, 62, 128, 223, 266, see also: ROC

empirical cross-entropy expected value of the posterior entropy, given a set of binary decision scores, in which the values of the scores are *integrated over the entire domain* [115].

Note: ECE plots are motivated from information theory as the divergence between labeled score sets to the ground-of-truth regarding the prior belief π . In APE plots, the integral of APE characteristics is C_{llr} . In ECE plots, C_{llr} can be read off as the ECE value at $\pi = 0$ (representing the generalized empirical cross-entropy). Other ECE values might be interpreted as C_{llr} values that are π dependently weighted between propositions \mathcal{A} and \mathcal{B} , 70, 127, 164, 247, see also: BDF & APE

enrolment *act of creating and storing a biometric enrolment data record in accordance with an enrolment policy* [1].

Note: act of creating and storing reference representation(s) in accordance with an operational policy, 29, 31, 91, 184

equal error rate *the value at which the Type I and Type II error proportions are equal* [95].

Note: EER estimates can be linear, based on theoretical score distribution assumptions or based on the ROCCH. In this dissertation, EERs are estimated as the ROCCH-EER, cf. [79], 45, 63, 113, 131, 168, 190, 217, 259, see also: FAR, FRR, FNMR, FMR, APCER, BPCER & ROCCH

expectation-maximization algorithm iterative parameter estimation, cf. [59]; statistical method to maximize the likelihood of a model. Regarding members of the exponential family distributions, their natural parameters are updated during maximization steps to derive expectational parameters, which are

used during the next expectation step to estimate the data fit of the updated model, [84](#)

failure-to-acquire rate *the proportion of acquisition processes for verification or identification attempts for which the system fails to capture or locate a sample of sufficient quality* [[25](#), 2019 CD1 revision].

In the 2016 re-confirmed standard [[25](#)] as: *proportion of verification or identification attempts for which the system fails to capture or locate an image or signal of sufficient quality*, [31](#)

failure-to-enrol rate *the proportion of acquisition processes for verification or identification attempts for which the system fails to capture or locate a sample of sufficient quality. [...] [25, 2019 CD1 revision].*

In the 2016 re-confirmed standard [[25](#)] as: *proportion of the population for whom the system fails to complete the enrolment process*, [31](#)

false accept rate *the proportion of transactions with false biometric claims erroneously accepted. A transaction may consist of one or more attempts depending on the decision policy. [...] [25, 2019 CD1 revision]:*

$$\text{FAR} = \text{FMR} \times (1 - \text{FTA}).$$

In the 2016 re-confirmed standard [[25](#)] as: *proportion of verification transactions with wrongful claims of identity that are incorrectly confirmed*, [32](#), [151](#)

false match rate *proportion of completed non-mated comparison trials in which the non-mated probe and reference are falsely declared “match”. [...] [25, 2019 CD1 revision].*

In the 2016 re-confirmed standard [[25](#)] as: *proportion of zero-effort impostor attempt samples falsely declared to match the compared non-self template.*

Note: speaker recognition and forensic science communities may refer to the *false alarm probability* and the *false acceptance probability*, respectively, [31](#), [58](#), [106](#), [118](#), [240](#)

false non-match rate *the proportion of completed mated comparison trials in which the mated probe and reference are falsely declared “non-match”. [...] [25, 2019 CD1 revision].*

In the 2016 re-confirmed standard [[25](#)] as: *proportion of genuine attempt samples falsely declared not to match the template of the same characteristic from the same user supplying the sample.*

Note: speaker recognition and forensic science communities may refer to the *miss probability* and the *false rejection probability*, respectively, [31](#), [58](#), [118](#), [240](#)

false reject rate *the proportion of verification transactions with true biometric claims erroneously rejected. A transaction may consist of one or more attempts depending on the decision policy. [...] [25, 2019 CD1 revision]:*

$$\text{FRR} = \text{FTA} + \text{FNMR} \times (1 - \text{FTA}).$$

In the 2016 re-confirmed standard [[25](#)] as: *proportion of verifi-*

cation transactions with truthful claims of identity that are incorrectly denied, 32, 151

FNMR at a 1% FMR security motivated theoretical error rate, depicting an indication to a preserved convenience level at a targeted level of security.

Note: FMR₁₀₀ requirements might denote $\text{FMR}_{100} \leq 5\%$, i.e., requiring the EER to be lower than 5%.

Note: in this dissertation, FMR_{100s} are estimated by the ROCCH, cf. ROCCH-EER in [79], 113, 160, 168, 191, 259, see also: FNMR, FMR, EER & ROCCH

function of quality estimate additive term in score calibration based on quality vectors for adaptively taking estimated quality changes of a certain reference–probe comparison into account (as an alternative to QMFs), 16, 195, 248, see also: sample quality, q-vector, UAC & QMF

Gaussian mixture model weighted sum of normal distributions for which the sum of the weights equals one [126]; utilizations: i) modeling of any multi-variate distribution, ii) statistic clustering, iii) data generation (emission of possible observations and its associated likelihood of observation), and iv) classification (class discrimination), 76, 167, 221, 249, 261

generalized false accept rate combination of enrolment, probe acquisition, and comparison errors regarding false accepts [25]:

$\text{GFAR} = (1 - \text{FTE})^2 \times \text{FAR}$ for scenario evaluation,

$\text{GFAR} = (1 - \text{FTE}) \times \text{FAR}$ for technology evaluation, 32

generalized false reject rate combination of enrolment, probe acquisition, and comparison errors regarding false rejects [25]:

$\text{GFRR} = \text{FTE} + (1 - \text{FTE} \times \text{FRR})$ for scenario and technology evaluation, 32

homomorphic encryption computation on encrypted data using operations that preserve algebraic structures of the plaintext data [344], 9, 227, 245, 249

identity concealer *subversive biometric capture subject who attempts to avoid being matched to their own biometric reference* [1], 35

impostor *subversive biometric capture subject who attempts to being matched to someone else's biometric reference.*

Note: Oxford defines impostor as a person who assumes a false identity in order to deceive or defraud [1], 35

indifferent biometric capture subject *biometric capture subject who is unconcerned with the achievement of a successful biometric acquisition process* [1]

intermediate-sized vector representation of acoustic space as a latent variable after factor analysis, mapping duration-variable audio samples into a fix-dimensional feature space, cf. [130], 17, 76, 114, 163, 183, 216, 247, 255

Kullback-Leibler divergence relative entropy; the divergence of two probability distributions, cf. [64], 69, 166, 190, 211

likelihood ratio ratio of proposition-conditional probabilities, where propositions are mutually exclusive and exhaustive in modeling a specific world.

Note: the terminology is currently under debate as the term *likelihood ratio* can be confused with the *likelihood ratio test*, which is not applicable to Bayesian inference. The term *Bayes factor* was proposed for likelihood ratios under a Bayesian t-test. However, rather than testing *hypotheses*, the [Bayesian decision framework](#) accommodates to formally denote thresholds for decision making on *propositions* based on prior and cost beliefs. As the terminology debate is not in the scope of this dissertation, the term is referred to as it is conventionally used in the speaker recognition community, 19, 24, 47, 127

log-likelihood ratio log-compressed LR.

Note: conventionally, the natural logarithm is utilized in the speaker recognition community, whereas the base 10 logarithm is preferred in the forensic science community as the score is intended to be reported in rather human-comprehensible scales, 5, 17, 31, 39, 48, 56, 77, 109, 115, 127, 173, 183, 215, 245, 263, *see also*: LR

mel-frequency cepstral coefficient sound representation [122, 123]; acoustic signals are processed by the short-term Fourier amplitude spectrum (for 20 to 30 ms time windows); powers of the spectrum are processed by a triangular filterbank mapped onto the mel scale [124] (motivated by a psycho-acoustic study on melodic perception); the inverse spectrum (*cepstral* from the anagram *cepstrum*) is obtained by the cosine transform of the log-compressed mel powers, resulting in the amplitude information (and omitting phase information), 17, 76, 109, 117, 174, 183, 218

minimum DCF minimum of the DCF. For fixed weights, a minimum is found by evaluating the DCF across all thresholds (equivalent to the DCF value after PAV), 60, 71, 118, 130, 240, 261, *see also*: DCF & PAV

minimum ECE minimum of the ECE; equivalent to the ECE value after PAV, 71, *see also*: ECE & PAV

non-subversive biometric capture subject *biometric capture subject who does not attempt to subvert the correct and intended system policy of the biometric capture subsystem* [1]

non-subversive user *user of a biometric system who does not attempt to subvert the correct and intended system policy* [1]

normalized Bayes error rate normalization of APE plots to enhance visual comparability; y-axis values are normalized by the default performance (that a coin tossing system would have) depending on the parameterization summarized by the x-axis value [28, 79], 61, 155, see also: APE

normalized ECE normalization of ECE plots to enhance visual comparability (motivated by NBER plots). Y-axis values are normalized by the default performance (that a coin tossing system would have) depending on the parameterization summarized by the x-axis value.

Note: in NECE plots, C_{llr} can be read off at $\pi = 0$ as the normalization denominator is one at $\pi = 0$, 15, 129, 247, see also: ECE & NBER

pool adjacent violators algorithm non-linear mapping of (uncalibrated) scores to (log-) likelihood ratios by isotonic regression [28, 78, 105, 106], 142, 211, 248, 263, see also: LR & LLR

presentation artefact *artificial object or representation presenting a copy of biometric characteristics or synthetic biometric patterns* [1, 43], 34, 217, 219

presentation attack *presentation to the biometric capture subsystem with the goal of interfering with the operation of the biometric system* [1, 43].

Note: depending on the evaluation type, the scope of presentation attacks is extended to also incorporate *modified biometric samples*, i.e., in *academic evaluation*, see [93]. The speaker recognition community synonymously uses the term *spoofing* for presentation attacks and the terms *physical access* and *logical access* for presentation attacks at the sensor level and as *modified biometric samples* [45], 2, 8, 34, 125, 215, 219, 251

presentation attack detection *automated determination of a presentation attack* [1], 8, 34, 46, 109, 113, 215, 249, see also: presentation attack

presentation attack instrument *biometric characteristic or object used in a presentation attack* [1, 43], 34, 125, 217, 249, see also: presentation attack

presentation attack instrument species *class of presentation attack instruments created using a common production method that is based on different biometric characteristics* [93], 218, see also: PAI

probabilistic linear discriminant analysis (binary) classifier examining latent feature subspaces regarding within and between class covariances to return (log-) likelihood ratio scores [60, 131], 10, 18, 76, 98, 111, 114, 163, 183, 227, 249, 255, see also: LR & LLR

quality measure function additive term during score calibration based on (proxy) quality measures taking into account ob-

servable quality changes of a certain reference–probe comparison (e.g., duration and SNR) (as an alternative to FQEs), [7](#), [193](#), [247](#), *see also*: [sample quality](#) & [FQE](#)

quality vector estimated posterior probabilities of (audio) *quality* conditions based on latent acoustic features (i-vectors), represented in vector form (motivated by UACs). For the purpose of speaker recognition, conditions are sampled according to well-studied quality impacts to recognition performance (targeting more detailed levels of certain quality types rather than broad levels ranging over a variety of audio characteristics), [16](#), [78](#), [109](#), [162](#), [184](#), [247](#), *see also*: [UAC](#) & [sample quality](#)

receiver operating characteristic trade-off (plot) between the true positive rate and the Type I error rate, [33](#), [62](#), [111](#), [127](#), [246](#), [266](#)

ROC's convex hull isotropic performance bound to the comparison of empirical Type I and Type II error rates in an orthogonal diagram of linear axes scale; the ROCCH is derived by i) the geometric convex hull and ii) by the PAV-LLR algorithm, mapping the DCF of well-calibrated scores onto the ROC space—the ROCCH is *where minDCF lives* [[79](#)], [62](#), [71](#), [111](#), [127](#), [240](#), [246](#), [266](#), *see also*: [minDCF](#), [PAV](#) & [ROC](#)

rule of 3 a rule of thumb (motivated by the Bernoulli distribution [[91](#)]). When observing zero errors on N sample points, the true error rate has an upper bound of $p \approx \frac{3}{N}$, holding in 95% of all samplings (confidence interval). For a 90% confidence interval, the rule of thumb refers to $p \approx \frac{2}{N}$.
Note: the rule of 3 is part of the *informative* annex of the standard [[25](#)], [33](#), [38](#), [41](#), [110](#)

rule of 30 a rule of thumb (motivated by the Bernoulli distribution using the Wald test, assuming a large amount of sampling points and very low error rates) for sustaining that the true error rate is bounded by a relative margin of $\pm 30\%$, at least 30 errors need to be observed. This hypothesis holds in 90% of all (experimental setup) samplings, i.e., with a 90% confidence interval [[92](#)].

Note: the rule of 30 is part of the *informative* annex of the standard [[25](#)], [33](#), [38](#), [41](#), [61](#), [110](#)

sample quality in the Concise Oxford English Dictionary (COED) [[86](#)] as: *the degree of excellence of something*. The COED further provides a phonetics definition: *distinguishing characteristic or characteristics of a speech sound*.

Note: the harmonized biometric vocabulary [[1](#)] defines sample quality as a predictor of biometric performance (to its biometric utility), since the biometrics standardization committee puts emphasis on black box testing rather than on the research and development of biometric systems (where white

- box testing is practiced). Regarding this term, this dissertation differs from ISO/IEC 2382-37 [1] by considering quality as a distinguishing characteristic or characteristics of a speech sound before and within acoustic feature processing, 27, 28, 183, 193
- signal-to-noise ratio** ratio of the power of the (speech) signal to the power of the noise signal, conventionally in dB, i.e., at a $10 \log_{10}$ compression, 7, 118, 172, 183, 248
- speaker recognition evaluation** technology evaluations for speaker recognition, e.g., the NIST evaluation series, 6, 58, 113, 174, 198, 251, 259, *see also*: NIST
- strictly proper scoring rule** a cost measurement of the accuracy of probabilistic predictions, where the minimal expected cost reflects the true set of probabilities; strictly proper scoring rules uniquely minimize the expected cost [28, 98], 39, 49, 72, 110, 154, 250, 266
- subject** biometric system data subject: *individual whose individualized biometric data is within the biometric system* [1], 6, 24, 67, 118, 163, 182, 217, 247, 249
- subversive biometric capture subject** *biometric capture subject who attempts to subvert the correct and intended policy of the biometric capture subsystem* [1]
- subversive user** *user of a biometric system who attempts to subvert the correct and intended system policy* [1], 217
- two covariance model** binary classifier which is formulatable as full subspace PLDA and as a primal support vector machine [133, 158], 10, 98, 102, 117, 227, 249, *see also*: PLDA
- uncooperative biometric capture subject** *biometric capture subject motivated to not achieve a successful completion of the biometric acquisition process* [1]
- unified audio characterization** estimation of posterior probabilities of audio characteristics.
Note: originally proposed in [38] regarding eight classes of characteristic audio properties, i.e., *clean telephone* or *microphone* sensors, *noisy* in 8 dB, 15 dB, 20 dB, *reverberated* speech with reverberation time factors 0.3, 0.5, 0.7, 7, 185, 247, *see also*: sample quality, FQE & q-vector
- universal background model** model representing/clustering the observed acoustic space [126].
Note: in traditional GMM-UBM systems, GMMs model certain subjects, where by contrast, UBMs (approximatively) model all other subjects, sometimes referred to as a *cohort model*. In conventional i-vector/PLDA systems, the UBM is basis to the extraction of probabilistic embeddings (i-vectors) [130], 76, 82, 114, 170, 228, 261, *see also*: GMM & i-vector

US National Institute of Standards and Technology measurements standards laboratory; non-regulatory agency of the US department of commerce, [6](#), [58](#), [113](#), [174](#), [198](#), [240](#), [251](#), [259](#)

user (of a biometric system) *any person or organization interacting in any way with a biometric system* [[1](#)], [27](#)

voice activity detection automated segmentation of biometric voice samples into *speech* and *non-speech* frames, cf. [[225](#)], [81](#), [118](#), [163](#), [171](#), [183](#), [247](#)

zero-effort impostor randomly selected pair of a non-mated reference and probe for the purpose of creating a non-mated comparison trial in (technology) performance evaluation, [33](#), see also: [impostor](#)

BIBLIOGRAPHY

- [1] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 2382-37:2017 information technology - vocabulary - part 37: Biometrics*, International Organization for Standardization, 2017.
- [2] C. Carolan, "Biometrics, interoperability and large-scale IT systems: The future of EU smart borders," in *Proc. GI/IEEE Intl. Conf. of the Biometrics Special Interest Group (BIOSIG)*, Keynote, 2017.
- [3] N. Connor and C. Wei, *Beijing park uses face recognition software to wipe out toilet paper theft*, The Telegraph, [Online] <http://www.telegraph.co.uk/news/2017/03/20/beijing-park-uses-face-recognition-software-wipe-toilet-paper>, accessed: 2017-10-05, Mar. 2017.
- [4] W. Grundzien, "Synopsis Biometrie: Update 2014," Bundesverband deutscher Banken, Tech. Rep., 2014.
- [5] K. Amino, A. Drygajlo, A. Eriksson, M. Kulshreshtha, J. H. L. Hansen, *et al.*, *Forensic Speaker Recognition*. Springer-Verlag New York, 2012.
- [6] M. Falcone, *The OCTAVE experience in the labs and on field — the way forward*, EU Octave Project, Final Workshop presentation, [Online] <https://www.octave-project.eu/wp-content/uploads/2017/07/14.30-Falcone-The-Way-Forward.pdf>, accessed: 2017-10-06, Jun. 2017.
- [7] M. Scuccimarra, *The OCTAVE experience on field — linat airport*, EU Octave Project, Final Workshop presentation, [Online] <https://www.octave-project.eu/wp-content/uploads/2017/07/14.40-Scuccimarra-SEA-Trial.pdf>, accessed: 2017-10-06, Jun. 2017.
- [8] P. Sawers, *Amazon wants Alexa everywhere, opens microphone tech behind echo to third-party devices*, Venture Beat, [Online] <https://venturebeat.com/2017/04/13/amazon-wants-alexa-everywhere-opens-microphone-tech-behind-echo-to-third-party-devices>, accessed: 2017-10-03, Apr. 2017.
- [9] N. Statt, *Amazon may give app developers access to Alexa audio recordings — a substantial shift in Amazon's stance on consumer privacy*, The Verge, [Online] <https://www.theverge.com/2017/7/12/15960596/amazon-alexa-echo-speaker-audio-recordings-developers-data>, accessed: 2017-10-03, Jul. 2017.

- [10] The IDIAP – TeSLA Team, *Speaker recognition in TeSLA*, Applied trust-based e-assessment (EU TeSLA project), [Online] <http://tesla-project.eu/speaker-recognition-tesla/>, accessed: 2017-10-05, Mar. 2017.
- [11] KnuVerse, *Voice authentication in real world noisy environments*, Speech Technology Magazine, White Paper, Jan. 2017.
- [12] Voxeo, *Fighting fraud without frustrating customers*, Speech Technology Magazine, White Paper, Nov. 2013.
- [13] D. Bushell-Embling, *HSBC adds voice authentication to digital banking in Hong Kong*, FINTECH innovation, [Online] <https://www.enterpriseinnovation.net/article/hsbc-adds-voice-authentication-digital-banking-hong-kong-1664173975>, accessed: 2017-10-03, Apr. 2017.
- [14] R. Amadeo, *My voice is my passport: Android gets a “Trusted Voice” smart lock — “OK Google” voice commands can get authorization from the sound of your voice*, Ars Technica, [Online] <https://arstechnica.com/gadgets/2015/04/my-voice-is-my-passport-android-gets-a-trusted-voice-smart-lock/>, accessed: 2017-10-03, Apr. 2015.
- [15] G. Heigold, I. Moreno, S. Bengio, and N. M. Shazeer, “End-to-end text-dependent speaker verification,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, [Online] <https://research.google.com/pubs/pub44681.html>, accessed: 2017-10-03, 2016, pp. 5115–5119.
- [16] S. Held, *Siri soll dank Stimmerkennung bald nur noch auf den Besitzer hören*, Mobil Branche Newsletter, [Online] <http://mobilbranche.de/2017/04/siri-stimmerkennung-personalisierung>, accessed: 2017-10-06, Apr. 2017.
- [17] Tagesschau, *Sprachassistent: Amazon wertet Alexa-Aufnahmen aus*, [Online] <https://www.tagesschau.de/wirtschaft/amazon-alexa-107.html>, accessed: 2019-04-25, Apr. 2019.
- [18] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, “The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding,” in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, Manuscript, 2019.
- [19] A. Nautsch, A. Jimenez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez-Barrero, D. Petrovska-Delcrétaz, G. Chollet, N. Evans, T. Schneider, J.-F. Bonastre, B. Raj, I. Trancoso, and C. Busch, “Preserving privacy in speaker and speech characterisation,” *Computer Speech and Language, Special issue on Speaker and language characterization and recognition: voice modeling, conversion,*

- synthesis and ethical aspects*, vol. 58, pp. 441–480, Nov. 2019, [Online] <https://doi.org/10.1016/j.csl.2019.06.001> [SurveyTalk] <https://www.youtube.com/watch?v=mywNMwZfbDo>, accessed 2019-10-20.
- [20] D. Meuwly and R. Veldhuis, “Forensic biometrics: From two communities to one discipline,” in *Proc. GI/IEEE Intl. Conf. of the Biometrics Special Interest Group (BIOSIG)*, 2012, pp. 207–218.
 - [21] J. Gonzalez-Rodriguez, “Evaluating automatic speaker recognition systems: An overview of the NIST speaker recognition evaluations (1996-2014),” *Loquens*, vol. 1, no. 1, e007, Jan. 2014, [Online] <http://loquens.revistas.csic.es/index.php/loquens/article/view/9/20>, accessed: 2017-08-03.
 - [22] National Institute of Standards and Technology (NIST), “NIST 2016 speaker recognition evaluation plan,” National Institute of Standards and Technology, Tech. Rep., 2016.
 - [23] S. E. Willis, L. Mc Kenna, S. Mc Dermott, A. Barrett, B. Rasmussen, *et al.*, *ENFSI guideline for evaluative reporting in forensic science*, [Online] http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf, accessed: 2017-05-22, European Network of Forensic Science Institutes, Mar. 2015.
 - [24] D. Meuwly, D. Ramos, and R. Haraksim, “A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation,” *Forensic Science International*, vol. 276, pp. 142–153, Jul. 2017.
 - [25] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 19795-1:2006. information technology – biometric performance testing and reporting – part 1: Principles and framework*, confirmed in 2011 and in 2016, International Organization for Standardization and International Electrotechnical Committee, Apr. 2006.
 - [26] N. Brümmer and J. du Preez, “Application-independent evaluation of speaker detection,” *Elsevier Computer Speech and Language (CSL)*, vol. 20, no. 2, pp. 230–275, Jul. 2006.
 - [27] D. Ramos-Castro, “Forensic evaluation of the evidence using automatic speaker recognition systems,” PhD thesis, Universidad Politécnica de Madrid, 2007.
 - [28] N. Brümmer, “Measuring, refining and calibrating speaker and language information extracted from speech,” PhD thesis, University of Stellenbosch, 2010.
 - [29] R. Haraksim, “Validation of likelihood ratio methods used for forensic evidence evaluation: Application in forensic fingerprints,” PhD thesis, University of Twente, 2014.
 - [30] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Communication*, vol. 52, no. 1, pp. 12–40, Jan. 2010.

- [31] J. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *Signal Processing Magazine, IEEE*, vol. 32, no. 6, pp. 74–99, Nov. 2015.
- [32] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 7663–7667.
- [33] National Institute of Standards and Technology (NIST), "The NIST year 2008 speaker recognition evaluation plan," National Institute of Standards and Technology, Tech. Rep., 2008.
- [34] —, "The NIST year 2010 speaker recognition evaluation plan," National Institute of Standards and Technology, Tech. Rep., 2010.
- [35] —, "The NIST year 2012 speaker recognition evaluation plan," National Institute of Standards and Technology, Tech. Rep., 2012.
- [36] —, "The 2013-2014 speaker recognition i-vector machine learning challenge," National Institute of Standards and Technology, Tech. Rep., 2014, pp. 1–2.
- [37] E. Khoury, B. Vesnicer, J. Franco-Pedroso, R. Violato, Z. Boulkenafet, *et al.*, "The 2013 speaker recognition evaluation in mobile environment," in *Proc. IEEE Intl. Conf. on Biometrics (ICB)*, 2013, pp. 2376–4201.
- [38] L. Ferrer, L. Burget, O. Plchot, and N. Scheffer, "A unified approach for audio characterization and its application to speaker recognition," in *Proc. Odyssey 2012: The Speaker and Language Recognition Workshop*, 2012, pp. 317–323.
- [39] M. I. Mandasari, "Speaker recognition systems in forensic conditions: The calibration and evaluation of the likelihood ratio," PhD thesis, Radboud Universiteit Nijmegen, 2017.
- [40] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot, and N. Scheffer, "Promoting robustness for speaker modeling in the community: The prism evaluation set," in *Proc. SRE11 Analysis Workshop*, 2011.
- [41] R. Saeidi, K. Lee, T. Kinnunen, *et al.*, "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2013.
- [42] N. Ratha, J. Connell, and R. Bolle, "Enhancing security and privacy of biometric-based authentication systems," *IBM Systems Journal*, vol. 40, no. 3, pp. 614–634, 2001.

- [43] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 30107-1. information technology - biometric presentation attack detection - part 1: Framework*, International Organization for Standardization, 2016.
- [44] ISO/IEC JTC1 SC27 Security Techniques, *ISO/IEC 24745:2011. information technology - security techniques - biometric information protection*, International Organization for Standardization, 2011.
- [45] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, et al., "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal on Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, Feb. 2017, [Online] <https://doi.org/10.1109/JSTSP.2017.2671435>, accessed: 2017-10-13.
- [46] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "ASVspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," ASVspoof consortium, Tech. Rep., 2014.
- [47] T. Kinnunen, N. Evans, J. Yamagishi, K. A. Lee, M. Sahidullah, M. Todisco, and H. Delgado, "ASVspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," ASVspoof consortium, Tech. Rep., 2016.
- [48] ASVspoof consortium, "ASVspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," ASVspoof consortium, Tech. Rep., 2018.
- [49] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. H. and M. Sahidullah, and A. Sizov, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2015, pp. 2037–2041.
- [50] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, [Online] http://www.isca-speech.org/archive/interspeech_2015/papers/i15_2037.pdf, accessed: 2017-08-07, 2017, pp. 2037–2041.
- [51] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, Manuscript, 2019.

- [52] T. Kinnunen, A. L. Kong, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Proc. Odyssey 2018: The Speaker and Language Recognition Workshop*, 2018, pp. 312–319.
- [53] S. Cumani, O. Plchot, and P. Laface, "Probabilistic linear discriminant analysis of i-vector posterior distributions," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 7644–7648.
- [54] N. Brümmer, A. Silnova, L. Burget, and T. Stafylakis, "Gaussian meta-embeddings for efficient scoring of a heavy-tailed PLDA model," in *Proc. Odyssey 2018: The Speaker and Language Recognition Workshop*, 2018, pp. 349–356.
- [55] European Council, *Regulation 2016/679 of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*, Apr. 2016.
- [56] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New York, NY, USA: John Wiley and Sons, 1973.
- [57] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, 2001.
- [58] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, 2006.
- [59] K. P. Murphy, *Machine Learning. A probabilistic perspective*. MIT Press, 2012.
- [60] S. J. D. Prince, *Computer Vision: Models, Learning and Inference*. Cambridge University Press, 2012.
- [61] D. van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker Classification I: Fundamentals, Features, and Methods*, ser. Lecture Notes in Computer Science, vol. 4343, [Online] <https://sites.google.com/site/nikobrummer/appindepeval-chapter.pdf>, accessed: 2017-06-20, Springer Berlin Heidelberg, 2007, pp. 330–353.
- [62] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, and T. Niemi, "Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition including guidance on the conduct of proficiency testing and collaborative exercises," *European Network of Forensic Science Institutes, Tech. Rep.*, 2015.
- [63] A. Papoulis and S. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed. McGraw-Hill, 2002.

- [64] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.
- [65] A. Nautsch, D. Ramos, and D. Meuwly, “Binary-decision error tradeoff (BET) plots,” *Manuscript*, 2019.
- [66] A. Nautsch, D. Meuwly, D. Ramos, J. Lindh, and C. Busch, “Making likelihood ratios digestible for cross-application performance assessment,” *IEEE Signal Processing Letters (SPL)*, vol. 24, no. 10, pp. 1552–1556, Oct. 2017, [Online] <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8025342> [Code] <https://doi.org/10.24433/C0.154591c8-9d3f-47eb-b656-3aff245fd5c1>, accessed: 2017-10-05.
- [67] —, “Making likelihood ratios digestible for cross-application performance assessment,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, SPL presentation, 2018.
- [68] A. Nautsch, R. Bamberger, and C. Busch, “Decision robustness of voice activity segmentation in unconstrained mobile speaker recognition environments,” in *Proc. GI/IEEE Intl. Conf. of the Biometrics Special Interest Group (BIOSIG)*, 2016, pp. 135–146.
- [69] A. Nautsch, C. Rathgeb, R. Saeidi, and C. Busch, “Entropy analysis of i-vector feature spaces in duration-sensitive speaker recognition,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 4674–4678.
- [70] A. Nautsch, R. Saeidi, C. Rathgeb, and C. Busch, “Analysis of mutual duration and noise effects in speaker recognition: Benefits of condition-matched cohort selection in score normalization,” in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2015, pp. 3006–3010.
- [71] —, “Robustness of quality-based score calibration of speaker recognition systems with respect to low-SNR and short-duration conditions,” in *Proc. Odyssey 2016: The Speaker and Language Recognition Workshop*, 2016, pp. 358–365.
- [72] A. Nautsch, S. T. Steen, and C. Busch, “Deep quality-informed score normalization for privacy-friendly speaker recognition in unconstrained environments,” in *Proc. GI/IEEE Intl. Conf. of the Biometrics Special Interest Group (BIOSIG)*, 2017, pp. 243–250.
- [73] U. Scherhag, A. Nautsch, C. Rathgeb, and C. Busch, “Unit-selection attack detection based on unfiltered frequency-domain features,” in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2016, pp. 2209–2213.

- [74] A. Nautsch, S. Isadskiy, J. Kolberg, M. Gomez-Barrero, and C. Busch, "Homomorphic encryption for speaker recognition: Protection of biometric templates and vendor model parameters," in *Proc. Odyssey 2018: The Speaker and Language Recognition Workshop*, 2018, pp. 16–23.
- [75] A. Nautsch, H. Hao, T. Stafylakis, C. Rathgeb, and C. Busch, "Towards PLDA-RBM based speaker recognition in mobile environment: Designing stacked/deep PLDA-RBM systems," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 5055–5059.
- [76] D. Ramos, R. Haraksim, and D. Meuwly, "Likelihood ratio data to report the validation of a forensic fingerprint evaluation method," *Data in Brief*, vol. 10, pp. 75–92, Feb. 2017, [Online] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5144651>, accessed: 2018-05-17.
- [77] R. Karbauskaite and M. D. Mansilla, "Best practice technical guidelines for automated border control (ABC) systems," FRONTEX, Tech. Rep., 2012, [Online] http://btn.frontex.europa.eu/system/files/private/resources/best_practice_technical_guidelines_for_abc_v2.0.pdf, accessed: 2018-04-23.
- [78] N. Brümmer and J. du Preez, *The PAV algorithm optimizes binary proper scoring rules*, 2009. eprint: [arXiv:1304.2331](https://arxiv.org/abs/1304.2331).
- [79] N. Brümmer and E. de Villiers, "The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing," AGNITIO Research, South Africa, Tech. Rep., Dec. 2011.
- [80] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, and J. Gonzalez-Rodriguez, "Deconstructing cross-entropy for probabilistic binary classifiers," *Entropy*, vol. 20, no. 3, p. 208, Mar. 2018.
- [81] D. Ramos, R. P. Krish, J. Fierrez, and D. Meuwly, "From biometric scores to forensic likelihood ratios," in *Handbook of Biometrics for Forensic Science*. Springer, 2017, ch. 14, pp. 305–327.
- [82] A. Nautsch, D. Ramos, J. González-Rodríguez, C. Rathgeb, and C. Busch, "Making decisions with biometric systems: The usefulness of a Bayesian perspective," in *Proc. NIST Intl. Biometric Performance Testing Conf. (IBPC)*, 2016.
- [83] D. Ramos and J. Gonzalez-Rodriguez, "Reliable support: Measuring calibration of likelihood ratios," *Forensic Science International*, vol. 230, pp. 156–169, 2013.
- [84] A. Nautsch and C. Busch, "Voice biometrics: How the technology is standardized," in *Voice Biometrics*, C. Garcia-Mateo and G. Chollet, Eds., Manuscript, IET, 2019.

- [85] J. L. Wayman, R. McIver, P. Wagget, S. Clarke, M. Mizoguchi, C. Busch, N. Delvaux, and A. Zudenzov, "Vocabulary harmonisation for biometrics: The development of ISO/IEC 2382 part 37," *IET Biometrics*, vol. 3, no. 1, pp. 1–8, Jun. 2014.
- [86] Oxford Dictionaries, *Concise Oxford English Dictionary*, 12th ed. Oxford University Press, Jul. 2011.
- [87] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC SC37 SD11 general biometric system*, International Organization for Standardization, May 2008.
- [88] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, 1997, pp. 1895–1898.
- [89] J. Navratil and G. Ramaswamy, "The awe and mystery of t-norm," in *Proc. ESCA Eur. Conf. on Speech Comm. and Tech. (EuroSpeech)*, 2003, pp. 2009–2012.
- [90] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.
- [91] B. Jovanovic and P. Levy, "A look at the rule of three," *The American Statistician*, pp. 137–139, 1997.
- [92] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective," *Elsevier Science Speech Communication*, vol. 31, pp. 225–254, Jun. 2000.
- [93] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 30107-3. information technology - biometric presentation attack detection - part 3: Testing and reporting*, International Organization for Standardization, 2017.
- [94] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and antispooofing in the i-vector space," *IEEE Trans. on Information Forensics and Security*, vol. 10, no. 4, pp. 821–832, Apr. 2015.
- [95] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC SC37 SD2 harmonized biometric vocabulary (v33)*, International Organization for Standardization, Jan. 2019.
- [96] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, Apr. 1950.
- [97] M. H. DeGroot and S. E. Fienberg, "The comparison and evaluation of forecasters," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 32, no. 1/2, pp. 12–22, 1983.

- [98] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007, [Online] <https://www.stat.washington.edu/raftery/Research/PDF/Gneiting2007jasa.pdf>, accessed: 2018-02-02.
- [99] N. Brümmer, L. Burget, P. Garcia, O. Plchot, J. Rhodin, *et al.*, "Meta-embeddings: A probabilistic generalization of embeddings in machine learning," JHU HLTCOE 2017 SCALE Workshop, Tech. Rep., 2017, [Online] <https://github.com/bsxfan/meta-embeddings/tree/master/theory>, accessed: 2017-12-08.
- [100] T. O'Hagan, "Dicing with the unknown," *Significance*, vol. 1, no. 3, pp. 132–133, Sep. 2004, [Online] http://www.stat.columbia.edu/~gelman/stuff_for_blog/ohagan.pdf, accessed: 2017-09-11.
- [101] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 19784-1:2006. information technology – biometric application programming interface – part 1: BioAPI specification*, International Organization for Standardization, Mar. 2006.
- [102] K. Shedden, *Binomial confidence intervals*, Department of Statistics, University of Michigan, [Online] http://dept.stat.lsa.umich.edu/~kshedden/Courses/Stat485/Notes/binomial_confidence_intervals.pdf, accessed: 2017-12-08, 2013.
- [103] H. Jeffreys, *Theory of Probability*, 3rd ed. Oxford Classic Texts in the Physical Sciences, 1961.
- [104] N. Brümmer and E. de Villiers, *What is the 'relevant population' in bayesian forensic inference?* 2014. eprint: [arXiv:1403.6008](https://arxiv.org/abs/1403.6008).
- [105] M. Ayer, H. Brunk, G. Ewing, W. Reid, and E. Silverman, "An empirical distribution function for sampling with incomplete information," *Ann. Math. Statist.*, vol. 26, no. 4, pp. 641–647, 1955, [Online] <https://projecteuclid.org/euclid.aoms/1177728423>, accessed: 2017-05-16.
- [106] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proc. of the Eighth Intl. Conf. on Knowledge Discovery and Data Mining*, 2002, pp. 694–699.
- [107] N. Brümmer, *Optimization of the accuracy and calibration of binary and multiclass pattern recognizers, for wide ranges of applications*, [Online] http://arantxa.ii.uam.es/~jms/seminarios_doctorado/abstracts2007-2008/20070226NBrummer.html, accessed: 2017-05-17, Feb. 2008.
- [108] T. Fawcett and A. Niculescu-Mizil, "PAV and the ROC convex hull," *Machine Learning*, vol. 68, no. 1, pp. 97–106, Jul. 2007.

- [109] R. Royall, "On the probability of observing misleading statistical evidence," *Journal of the American Statistical Assoc.*, vol. 95, no. 451, pp. 760–768, 2000.
- [110] J. J. Lucena-Molina, D. Ramos-Castro, and J. Gonzalez-Rodriguez, "Performance of likelihood ratios considering bounds on the probability of observing misleading evidence," *Law, Probability and Risk*, vol. 14, no. 3, pp. 175–192, 2015, [Online] <https://doi.org/10.1093/lpr/mgu022>, accessed: 2017-11-29.
- [111] T. Dunstone and N. Yager, *Biometric System and Data Analysis*. Springer, 2009.
- [112] A. Alexander, O. Forth, J. Nash, and N. Yager, "Zooplots for speaker recognition with tall and fat animals," in *Proc. Intl. Assoc. for Forensic Phonetics and Acoustics (IAFPA)*, [Online] <http://www.oxfordwaveresearch.com/papers/Alexander-Forth-Nash-Yager-IAFPA-2014-Abstract.pdf>, accessed: 2017-11-29, 2014.
- [113] R. Giot, R. Bourqui, and M. El-Abed, "Zoo graph: A new visualisation for biometric system evaluation," in *Proc. IEEE Intl. Conf. on Information Visualisation (IV)*, [Online] <https://hal.archives-ouvertes.fr/hal-01355690>, accessed: 2017-11-29, 2016, pp. 190–195.
- [114] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "SHEEP, GOATS, LAMBS and WOLVES — a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," in *Proc. Intl. Conf. on Spoken Language Processing (ICSLP)*, 1998, pp. 1351–1354.
- [115] D. Ramos and J. Gonzalez-Rodrigues, "Cross-entropy analysis of the information in forensic speaker recognition," in *Proc. IEEE Odyssey*, 2008.
- [116] N. Brümmer and G. Doddington, *Likelihood-ratio calibration using prior-weighted proper scoring rules*, 2013. eprint: [arXiv:1307.7981](https://arxiv.org/abs/1307.7981).
- [117] A. Nautsch, A. Lozano Diez, and D. Ramos, "Full/posterior PLDA in speaker recognition: Technical literature review," Hochschule Darmstadt, Universidad Autónoma de Madrid, Tech. Rep., 2016.
- [118] A. Treiber, A. Nautsch, J. Kolberg, T. Schneider, and C. Busch, "Privacy-preserving PLDA speaker verification using outsourced secure computation," *Speech Communication*, vol. 114, pp. 60–71, Nov. 2019, [Online] <https://doi.org/10.1016/j.specom.2019.09.004>, accessed 2019-10-16.

- [119] A. V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-16, no. 2, pp. 221–226, Jun. 1968.
- [120] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. Cambridge University, 2009.
- [121] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Hilton Waikoloa Village, Big Island, Hawaii, US, 2011.
- [122] J. S. Bridle and M. D. Brown, "An experimental automatic word-recognition system," Joint Speech Research Unit, Ruislip, England, JSRU Report 1003, 1974.
- [123] S. Davis and P. Mermelstein, "Comparison of parametric representations for mono-syllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing (ASSP)*, vol. 28, no. 4, pp. 357–366, 1980.
- [124] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America (JASA)*, vol. 8, no. 3, pp. 185–190, 1937.
- [125] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [126] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Conversational Speech, Digital Signal Processing*, vol. 10, pp. 19–41, 2000.
- [127] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," CRIM, Montreal, Tech. Rep. CRIM-06/08-13, 2005.
- [128] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 15, no. 4, pp. 1435–1447, May 2007.
- [129] O. Glembek, L. Burget, N. Dehak, N. Bümmer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 4057–4060.

- [130] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 19, no. 4, pp. 788–798, 2011.
- [131] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE Intl. Conf. on Computer Vision (ICCV)*, CVF, 2007.
- [132] D. Garcia-Romero and C. Epsy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2011, pp. 249–252.
- [133] S. Cumani, N. Brümmer, L. Burget, P. Laface, O. Plchot, and V. Vasilakakis, "Pairwise discriminative speaker verification in the i-vector space," *IEEE Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 21, no. 6, pp. 1217–1227, 2013.
- [134] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 165–170.
- [135] —, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2017, pp. 999–1003.
- [136] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.
- [137] S. Cumani, "Fast scoring of full posterior PLDA models," *IEEE/ACM Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 23, no. 11, pp. 2036–2045, 2015.
- [138] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, Aug. 1971.
- [139] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.
- [140] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1980.
- [141] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992, pp. 121–124.

- [142] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing (TSAP)*, vol. 2, no. 4, pp. 578–589, Oct. 1994.
- [143] O. Glembek, P. Matejka, O. Plchot, J. Pesan, L. Burget, and P. Schwarz, "Migrating i-vectors between speaker recognition systems using regression neural networks," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2015, pp. 2327–2331.
- [144] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proc. Odyssey 2016: The Speaker and Language Recognition Workshop*, 2016, pp. 283–290.
- [145] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey 2001: The Speaker and Language Recognition Workshop*, 2001, pp. 213–218.
- [146] A. Larcher, K. Lee, and S. Meignier, "An extensible speaker identification SIDEKIT in Python," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, [Online] <http://lium.univ-lemans.fr/sidekit>, accessed: 2017-05-15, 2016, pp. 5095–5099.
- [147] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006, pp. 97–100.
- [148] A. O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH) 2006, International Conference on Spoken Language Processing (ICSLP)*, 2006, pp. 1471–1474.
- [149] P. Kenny, T. Stafylakis, J. Alam, and M. Kockmann, "An i-vector backend for speaker verification," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2015, pp. 2307–2311.
- [150] P. A. Torres-Carrasquillo, F. Richardson, S. Nercessian, D. Sturim, W. Campbell, *et al.*, "The MIT-LL, JHU and LRDE NIST 2016 speaker recognition evaluation system," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2017, pp. 1333–1337.
- [151] R. Saeidi and P. Alku, "Accounting for uncertainty of i-vectors in speaker recognition using uncertainty propagation and modified imputation," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2015, pp. 3546–3550.

- [152] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2016, pp. 3434–3438.
- [153] S. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-end attention based text-dependent speaker verification," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 171–178.
- [154] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Conf. on Neural Information Processing Systems (NIPS)*, 2017, pp. 6000–6010.
- [155] S. Novoselov, A. Shulipa, I. Kremnev, A. Kozlov, and V. Schemelinin, "On deep speaker embeddings for text-independent speaker recognition," in *Proc. Odyssey 2018: The Speaker and Language Recognition Workshop*, 2018, pp. 378–385.
- [156] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, [Online] http://openaccess.thecvf.com/content_cvpr_2017/papers/Liu_SphereFace_Deep_Hypersphere_CVPR_2017_paper.pdf, accessed: 2019-01-21, 2017, pp. 212–220.
- [157] P.-M. Bousquet, J.-F. Bonastre, and D. Matrouf, "Identify the benefits of the different steps in an i-vector based speaker," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP): Proc. Iberoamerican Congress on Pattern Recognition*, Springer-Verlag Berlin Heidelberg, 2013, pp. 278–285.
- [158] S. Cumani and P. Laface, "Generative pairwise models for speaker recognition," in *Proc. Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014, pp. 273–279.
- [159] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey 2010: The Speaker and Language Recognition Workshop*, 2010.
- [160] A. Sizov, K. A. Lee, and T. Kinnunen, "Unifying probabilistic linear discriminant analysis variants in biometric authentication," in *Structural, Syntactic, and Statistical Pattern Recognition: Proc. Joint IAPR Intl. Workshop on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition (S+SSPR)*, Springer Berlin Heidelberg, 2014, pp. 464–475.
- [161] N. Brümmer and E. de Villiers, "The speaker partitioning problem," in *Proc. Odyssey 2010: The Speaker and Language Recognition Workshop*, 2010.

- [162] D. E. Sturim and D. A. Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker verification," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, pp. 741–744.
- [163] S. Cumani, P. D. Batzu, D. Colibro, C. Vair, P. Laface, and V. Vasilakakis, "Comparison of speaker recognition approaches for real applications," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2011, pp. 2365–2368.
- [164] D. Colibro, C. Vair, K. Farrell, N. Krause, G. Karvitsky, S. Cumani, and P. Laface, "Nuance - Politecnico di Torino (NPT) system description for NIST 2012 speaker recognition evaluation," in *Proc. NIST SRE'12 workshop*, 2012.
- [165] N. Brümmer, D. van Leeuwen, and A. Swart, "A comparison of linear and non-linear calibrations for speaker recognition," in *Proc. Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014, pp. 14–18.
- [166] M. I. Mandasari, R. Saeidi, M. McLaren, and D. A. van Leeuwen, "Quality measure functions for calibration of speaker recognition systems in various duration conditions," *IEEE Trans. on Audio, Speech and Language Processing (TASLP)*, vol. 21, no. 11, pp. 2425–2438, Nov. 2013.
- [167] M. I. Mandasari, R. Saeidi, and D. A. van Leeuwen, "Quality measures based calibration with duration and noise dependency for speaker recognition," *Speech Communication*, vol. 72, pp. 126–137, Sep. 2015.
- [168] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, *et al.*, "Bi-modal person recognition on a mobile phone: Using mobile phone data," in *Proc. IEEE ICME Workshop on Hot Topics in Mobile Multimedia*, 2012, pp. 635–640.
- [169] R. Saeidi and D. van Leeuwen, "The Radboud University Nijmegen submission to NIST SRE-2012," in *Proc. NIST SRE workshop*, 2012.
- [170] K. Lee, V. Hautamäki, T. Kinnunen, A. Larcher, C. Zhang, A. Nautsch, T. Stafylakis, *et al.*, "The I4U mega fusion and collaboration for NIST speaker recognition evaluation 2016," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2017, pp. 1328–1332.
- [171] K. A. Lee, V. Hautamaki, T. Kinnunen, H. Yamamoto, K. Okabe, V. Vestman, J. Huang, G. Ding, H. Sun, A. Larcher, R. K. Das, H. Li, M. Rouvier, P.-M. Bousquet, W. Rao, Q. Wang, C. Zhang, F. Bahmaninezhad, H. Delgado, J. Patino, Q. Wang, L. Guo, T. Koshinaka, J. Zhang, K. Shinoda, T. N. Trong, M. Sahidullah, F. Lu, Y. Tang, M. Tu, K. K. Teh, H. D. Tran, K. K.

- George, I. Kukanov, F. Desnoux, J. Yang, E. Yilmaz, L. Xu, J.-F. Bonastre, C. Xu, Z. H. Lim, E. S. Chng, S. Ranjan, J. H. L. Hansen, M. Todisco, and N. Evans, *I4U submission to NIST SRE 2018: Leveraging from a decade of shared experiences*, 2019. eprint: [arXiv:1904.07386](https://arxiv.org/abs/1904.07386).
- [172] D. Schnelle-Walka, S. Radeck-Arneth, C. Biemann, and S. Radomski, "An open source corpus and recording software for distant speech recognition with the Microsoft Kinect," in *Proc. ITG Symposium on Speech Communication*, 2014.
- [173] M. I. Mandasari, R. Saeidi, and D. A. van Leeuwen, "Calibration based on duration quality measure function in noise robust speaker recognition for NIST SRE'12," in *Proc. Biometric Technologies in Forensic Science (BTFS)*, 2013, pp. 1–5.
- [174] D. P. W. Ellis, *PLP and RASTA (and MFCC, and inversion) in Matlab*, [Online] <http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat>, accessed: 2013-10-10, 2005.
- [175] O. Glembek, *Joint factor analysis Matlab demo*, [Online] <http://speech.fit.vutbr.cz/software/joint-factor-analysis-matlab-demo>, accessed 2013-10-10, 2009.
- [176] D. van Leeuwen, A. Martin, M. Przybicki, and J. Bouten, "The NIST 2004 and TNO/NFI speaker recognition evaluations," *Elsevier Computer Speech and Language (CSL)*, vol. 20, no. 2-3, pp. 128–158, Aug. 2006.
- [177] C. Cieri, D. Miller, and K. Walker, "From Switchboard to Fisher: Telephone collection protocols, their uses and yields," in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, 2003, pp. 1597–1600.
- [178] J. Ortega-Garcia, J. Gonzalez-Rodriguez, and V. Marrero-Aguiar, "AHUMADA: A large speech corpus in spanish for speaker characterization and identification," *Speech Communication*, vol. 31, no. 2-3, pp. 255–264, Jun. 2000.
- [179] H. Nakasone and S. D. Beck, "Forensic automatic speaker recognition," in *Proc. 2001: A Speaker Odyssey – The Speaker Recognition Workshop*, 2001.
- [180] C. Cieri, J. P. Campbell, H. Nakasone, and D. Miller, "The Mixer corpus of multilingual, multichannel speaker recognition data," in *Proc. Intl. Conf. on Language Resources and Evaluation (LREC)*, [Online] <https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/lrec2004-mixer-corpus.pdf>, accessed: 2017-08-03, 2004, pp. 627–630.

- [181] C. Cieri, W. Andrews, J. P. Campbell, G. Doddington, J. Godfrey, *et al.*, "The Mixer and transcript reading corpora: Resources for multilingual, crosschannel speaker recognition research," in *Proc. Intl. Conf. on Language Resources and Evaluation (LREC)*, [Online] <https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/lrec2006-mixer-and-transcript-reading-corpora.pdf>, accessed: 2017-08-03, 2006, pp. 117–120.
- [182] C. Cieri, L. Corson, D. Graff, and K. Walker, "Resources for new research directions in speaker recognition: The Mixer 3, 4 and 5 corpora," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, [Online] <https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/interspeech2007-resources-for-mixer-3-4-5.pdf>, accessed: 2017-08-03, 2007, pp. 950–953.
- [183] L. Brandshain, C. Cieri, D. Graff, A. Neely, and K. Walker, "Speaker recognition: Building the Mixer 4 and 5 corpora," in *Proc. Intl. Conf. on Language Resources and Evaluation (LREC)*, [Online] <https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/lrec2008-speaker-recognition-mixer4-mixer5.pdf>, accessed: 2017-08-03, 2008, pp. 3551–3554.
- [184] L. Brandshain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "The Mixer 6 corpus: Resources for cross-channel and text independent speaker recognition," in *Proc. Intl. Conf. on Language Resources and Evaluation (LREC)*, [Online] http://www.lrec-conf.org/proceedings/lrec2010/pdf/792_Paper.pdf, accessed: 2017-08-04, 2010, pp. 2441–2444.
- [185] L. Brandschain, D. Graff, C. Cieri, K. Walker, and C. Caruso, "Greybeard - voice and aging," in *Proc. Intl. Conf. on Language Resources and Evaluation (LREC)*, 2010, pp. 2437–2440.
- [186] C. S. Greenberg, V. M. Stanford, A. F. Marting, M. Yadagiri, G. R. Doddington, *et al.*, "The 2012 NIST speaker recognition evaluation," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2013, pp. 1971–1975.
- [187] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text," in *Proc. Intl. Conf. on Language Resources and Evaluation (LREC)*, 2004, pp. 69–71.
- [188] National Institute of Standards and Technology (NIST), "The NIST year 2004 speaker recognition evaluation plan," National Institute of Standards and Technology, Tech. Rep., 2004.
- [189] —, "The NIST year 2005 speaker recognition evaluation plan," National Institute of Standards and Technology, Tech. Rep., 2005.

- [190] —, “The NIST year 2006 speaker recognition evaluation plan,” National Institute of Standards and Technology, Tech. Rep., 2006.
- [191] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 19795-1:2006. information technology – biometric performance testing and reporting – part 1: Principles and framework*, International Organization for Standardization and International Electrotechnical Committee, Apr. 2006.
- [192] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, “Temporally weighted linear prediction features for tackling additive noise in speaker verification,” *IEEE Signal Processing Letters (SPL)*, vol. 17, no. 6, pp. 599–602, Jun. 2010.
- [193] R. McAulay and M. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. on Acoustics, Speech and Signal Processing (ASSP)*, vol. 28, no. 2, pp. 137–145, 1980.
- [194] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, “Multi-condition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 4257–4260.
- [195] D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, C. S. Greenberg, *et al.*, “Summary and initial results of the 2013-2014 speaker recognition i-vector machine learning challenge,” in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2014, pp. 368–372.
- [196] I. W. Evett and P. Gill, “A discussion of the robustness of methods for assessing the evidential value of DNA single locus profiles in crime investigations,” *Electrophoresis*, vol. 12, no. 2-3, pp. 226–230, 1991.
- [197] I. W. Evett, G. Jackson, J. A. Lambert, and S. McCrossan, “The impact of the principles of evidence interpretation on the structure and content of statements,” *Science & Justice*, vol. 40, no. 4, pp. 233–239, 2000.
- [198] Association of Forensic Science Providers, “Standards for the formulation of evaluative forensic science expert opinion,” *Science and Justice*, vol. 49, no. 3, pp. 161–164, Sep. 2009, [Online] <http://dx.doi.org/10.1016/j.scijus.2009.07.004>, accessed: 2017-05-22.
- [199] A. Nordgaard, R. Ansell, W. Drotz, and L. Jaeger, “Scale of conclusions for the value of evidence,” *Law, Probability and Risk*, vol. 11, no. 1, pp. 1–24, 2012, [Online] <https://academic.oup.com/lpr/article-lookup/doi/10.1093/lpr/mgr020>, accessed: 2017-05-22.

- [200] D. H. Kaye, "The weight of evidence in law, statistics, and forensic science," in *Proc. NIST TC on Quantifying the Weight of Forensic Evidence*, [Online] https://www.nist.gov/sites/default/files/documents/2016/12/07/03_kaye_16-nist-woe-linear.pdf, accessed: 2017-05-22, 2016.
- [201] W. C. Thompson, "Lay reactions to quantitative statements about the weight of forensic science evidence," in *Proc. NIST TC on Quantifying the Weight of Forensic Evidence*, [Online] https://www.nist.gov/sites/default/files/documents/2016/12/07/09_thompson_nist_quant_conf.pdf, accessed: 2017-11-22, 2016.
- [202] G. S. Morrison and E. Enzinger, "A new paradigm for the evaluation of forensic evidence — and its implementation in forensic voice comparison," in *Proc. NIST TC on Quantifying the Weight of Forensic Evidence*, [Online] https://www.nist.gov/sites/default/files/documents/2016/12/07/04_morrisonenzinger_nist_workshop_2016_04_30a_optimized.pdf, accessed: 2017-11-22, 2016.
- [203] C. Drummond and R. C. Holte, "What ROC curves can't do (and cost curves can)," in *Proc. ROC Analysis in Artificial Intelligence (ROCAI)*, 2004.
- [204] J. Hernández-Orallo, P. Flach, and C. Ferri, "A unified view of performance metrics: Translating threshold choice into expected classification loss," *Journal of Machine Learning Research*, vol. 13, pp. 2813–2869, Oct. 2012.
- [205] R. Haraksim, D. Ramos, D. Meuwly, and C. E. H. Berger, "Measuring coherence of computer-assisted likelihood ratio methods," *Forensic Science International*, vol. 249, pp. 123–132, Apr. 2015.
- [206] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the ROC curve," *Machine Learning*, vol. 77, pp. 103–123, Jun. 2009.
- [207] D. J. Hand and C. Anagnostopoulos, "When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance?" *Pattern Recognition Letters*, vol. 34, no. 5, pp. 492–495, Apr. 2013.
- [208] J. Fiscus, *NIST SCLITE scoring package version 1.5*, [Online] <http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm>, accessed: 2016-02-26, NIST, 1998.
- [209] D. Yu, J. Li, and L. Deng, "Calibration of confidence measures in speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 19, no. 8, pp. 2461–2473, 2011.
- [210] P. A. Flach, "ROC analysis," in *Encyclopedia of Machine Learning and Data Mining*, Springer, Boston, MA, 2016.

- [211] A. Adler, R. Youmaran, and S. Loyka, "Towards a measure of biometric information," *Springer Pattern Analysis and Applications (PAA)*, vol. 12, no. 3, pp. 261–270, Sep. 2009.
- [212] J. Daugman, "Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1927–1935, Nov. 2006.
- [213] W. E. Burr, D. F. Dodson, and W. T. Polk, "Electronic authentication guideline, recommendations of the national institute of standards and technology, information security," National Institute of Standards and Technology (NIST), Tech. Rep., 2006.
- [214] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*. Springer, 2007.
- [215] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 29794-1:2009 information technology - biometric sample quality - part 1: Framework*, International Organization for Standardization, 2009.
- [216] R. Bamberger, "Sprachaktivitätserkennung in der Sprechererkennung," Master's thesis, Hochschule Darmstadt, 2015.
- [217] ITU-T, "Recommendation ITU-T P. 52 (1993), volume meters," Telecommunication Standardization Sector of ITU, Geneva, Switzerland, ITU-T P.56, 1994.
- [218] M. Brookes, *VOICEBOX: Speech processing toolbox for MATLAB*, Department of Electrical & Electronic Engineering, Imperial College, London, 2005.
- [219] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters (SPL)*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
- [220] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing (TASP)*, vol. 9, no. 5, pp. 504–512, 2001.
- [221] M. H. Moattar and M. M. Homayounpour, "A simple but efficient real-time voice activity detection algorithm," in *Proc. European Signal Processing Conference (EURASIP)*, 2009, pp. 2549–2553.
- [222] Z. H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 798–807, 2010.
- [223] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 7229–7233.

- [224] M. J. Alam, P. Kenny, P. Ouellet, T. Stafylakis, and P. Dumouche, "Supervised/unsupervised voice activity detectors for text-dependent speaker recognition on the RSR2015 corpus," in *Proc. Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014, pp. 123–130.
- [225] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 14, no. 2, pp. 412–424, 2006.
- [226] J. G. Gruber and L. Strawczynski, "Subjective effects of variable delay and speech clipping in dynamically managed voice systems," *IEEE Trans. on Communications*, vol. 33, no. 8, pp. 801–808, 1985.
- [227] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, "The voice activity detector for the Pan-European digital cellular mobile telephone service," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1989, pp. 369–372.
- [228] J. Kola, C. Epsy-Wilson, and T. Pruthi, "Voice activity detection," University of Maryland, Department of Electrical & Computer Engineering, College Park, Maryland, USA, MERIT BIEN final report, 2011.
- [229] M. Sahidullah and G. Saha, *Comparison of speech activity detection techniques for speaker recognition*, 2012. eprint: [arXiv:1210.0297](https://arxiv.org/abs/1210.0297).
- [230] M. McLaren and M. Graciarena, "softSAD: Integrated frame-based speech confidence for speaker recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 4694–4698.
- [231] H. Zhivomirov and P. Baranski, *Pink, red, blue and violet noise generation with Matlab implementation*, [Online] <http://www.mathworks.com/matlabcentral/fileexchange/42919-pink-red-blue-and-violet-noise-generation-with-matlab-implementation>, accessed: 2016-05-26, 2014.
- [232] J. Lyons, *Matlab speech feature generation scripts used internally in the signal processing lab at Griffith University*, [Online] https://github.com/jameslyons/spl_featgen, accessed: 2016-05-26, 2015.
- [233] P. Bak, C. Tang, and K. Wiesenfeld, "Self-organized criticality: An explanation of $1/f$ noise," *APS Physical Review Letters*, vol. 59, no. 4, pp. 381–384, 1987.
- [234] D. B. Dean, S. Sridharam, R. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2010, pp. 3110–3113.

- [235] A. Nautsch, C. Rathgeb, C. Busch, H. Reininger, and K. Kasper, "Towards duration invariance of i-vector-based adaptive score normalization," in *Proc. Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014, pp. 60–67.
- [236] H. Aronowitz, "Compensating inter-dataset variability in PLDA hyper-parameters for robust speaker recognition," in *Proc. Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014, pp. 280–286.
- [237] P. Rajan, T. Kinnunen, and V. Hautamäki, "Effect of multicondition training on i-vector PLDA configurations for speaker recognition," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2013, pp. 3694–3697.
- [238] S. O. Sadjadi, T. Hasan, and J. H. L. Hansen, "Mean Hilbert envelope coefficients (MHEC) for robust speaker recognition," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2012, pp. 1696–1699.
- [239] Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using vector Taylor series for speaker recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 6788–6791.
- [240] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting Baum-Welch statistics for speaker recognition," in *Proc. Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014, pp. 293–298.
- [241] R. Saeidi, R. F. Astudillo, and D. Kolossa, "Uncertain LDA: Including observation uncertainties in discriminative transforms," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 38, no. 7, pp. 1479–1488, 2015.
- [242] M. McLaren, A. Lawson, L. Ferrer, N. Scheffer, and Y. Lei, "Trial-based calibration for speaker recognition in unseen conditions," in *Proc. Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014, pp. 19–25.
- [243] H. Hotelling, "The generalization of student's ratio," *Annals of Mathematical Statistics*, vol. 2, no. 3, pp. 360–378, Aug. 1931.
- [244] A. Trujillo-Ortiz and R. Hernandez-Walls, *HotellingT2: Hotelling T-squared testing procedures for multivariate tests. A MATLAB file*. [Online] <http://www.mathworks.com/matlabcentral/fileexchange/2844-hotellingt2>, accessed: 2019-05-25, 2002.
- [245] D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "On the use of quality measures for text-independent speaker recognition," in *Proc. Odyssey 2004: The Speaker and Language Recognition Workshop*, 2004, pp. 105–110.

- [246] G. Aradilla, J. Vepa, and H. Bourlard, "Using posterior-based features in template matching for speech recognition," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2006, pp. 2570–2573.
- [247] S. T. Steen, "Deep learning for speaker recognition in noisy environments," Master's thesis, Danmarks Tekniske Universitet, 2017.
- [248] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Intl. Conf. on Computer Vision (ICCV)*, 2015, pp. 1026–1034.
- [249] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
- [250] D. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2014. eprint: [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [251] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting,," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [252] S. Ioffe and C. Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, 2015. eprint: [arXiv:1502.03167](https://arxiv.org/abs/1502.03167).
- [253] European Parliament and European Council, *Directive 2015/2366 of the European Parliament and of the Council of 25 November 2015 on payment services in the internal market*, Nov. 2015.
- [254] M. Faúndez-Zanuy, "On the vulnerability of biometric security systems," *IEEE Aerospace and Electronic Systems Magazine*, vol. 19, no. 6, pp. 3–8, 2004.
- [255] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, Feb. 2014.
- [256] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *Proc. Fala Workshop*, 2010, pp. 131–134.
- [257] C. Yang, G. Hammouri, and B. Sunar, "Voice passwords revisited," in *Proc. ICETE Intl. Conf. on Security and Cryptography (SECRYPT)*, 2012, pp. 163–171.
- [258] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *Proc. Intl. Carnahan Conference on Security Technology (ICCST)*, Oct. 2011, pp. 1–8.

- [259] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A.-M. Laukkanen, "I-vectors meet imitators: On vulnerability of speaker verification systems against voice mimicry," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2013, pp. 930–934.
- [260] U. J. Scherhag, "Presentation attack detection for state-of-the-art speaker recognition systems," Master's thesis, Hochschule Darmstadt, 2016.
- [261] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge," in *Proc. Annual Conference of the International Speech Communication Association, (INTERSPEECH)*, 2015, pp. 2052–2056.
- [262] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2015, pp. 2092–2096.
- [263] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection from a feature representation perspective," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 2119–2123.
- [264] C. Zhang, S. Ranjan, M. K. Nandwana, Q. Zhang, A. Misra, G. Liu, F. Kelly, and J. H. Hansen, "Joint information from nonlinear and linear features for spoofing detection: An i-vector/DNN based approach," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 1689–1699.
- [265] DFKI GmbH, *Unit selection voice creation and explanation on individual voice import components*, [Online] <https://github.com/marytts/marytts/wiki/UnitSelectionVoiceCreation>, accessed: 2016-02-22, 2014.
- [266] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2015, pp. 2062–2066.
- [267] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 19, no. 6, pp. 1791–1801, 2011.
- [268] Q. Li, "An auditory-based transform for audio signal processing," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 181–184.

- [269] T. B. Patel and H. A. Patil, "Effectiveness of fundamental frequency and strength of excitation for spoofed speech detection," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 5105–5109.
- [270] M. Wester, Z. Wu, and J. Yamagishi, "Human vs machine spoofing detection on wideband and narrowband data," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2015, pp. 2047–2051.
- [271] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 11, no. 7, pp. 674–693, Jul. 1989.
- [272] European Council, *Directive 2016/680 of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA*, Apr. 2016.
- [273] NIST Special Publication 500-290 ANSI/NIST-ITL 1-2011:Update 2015, *Data format for the interchange of fingerprint, facial and other biometric information*, National Institute of Standards and Technology, Aug. 2016.
- [274] C. Vaquero and P. Rodríguez, "On the need of template protection for voice authentication," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2015, pp. 219–223.
- [275] R. Cappelli, D. Maio, A. Lumini, and D. Maltoni, "Fingerprint image reconstruction from standard templates," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1489–1503, 2007.
- [276] J. Galbally, A. Ross, M. Gomez-Barrero, J. Fierrez, and J. Ortega-Garcia, "Iris image reconstruction from binary templates: An efficient probabilistic approach based on genetic algorithms," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1512–1525, 2013.
- [277] E. Moyakine, C. Colonnello, J. Butler, and C. Jasserand, *Discussion panel: SIIP and INGRESS research projects: Developing effective and sustainable biometric systems with a global reach*, EAB Research Projects Conference, [Online] https://www.eab.org/upload/documents/1279/08-eabrpc2017_SIIP_INGRESS.zip?ts=1517990107115, accessed: 2018-02-07, 2017.
- [278] S. Isadskiy, "Biometric information protection in i-vector feature space," Master's thesis, Hochschule Darmstadt, 2018.

- [279] C. Rathgeb and A. Uhl, "A survey on biometric cryptosystems and cancelable biometrics," *EURASIP Journal on Information Security*, vol. 2011, p. 3, Dec. 2011.
- [280] V. M. Patel, N. Ratha, and R. Chellappa, "Cancelable biometrics: A review," *IEEE Signal Proc. Magazine*, vol. 32, no. 5, pp. 54–65, 2015.
- [281] P. Campisi, Ed., *Security and Privacy in Biometrics*. Springer, 2013.
- [282] C. Aguilar-Melchor, S. Fau, C. Fontaine, G. Gogniat, and R. Sirdey, "Recent advances in homomorphic encryption: A possible future for signal processing in the encrypted domain," *IEEE Signal Processing Magazine*, vol. 30, no. 2, pp. 108–117, 2013.
- [283] M. Gomez-Barrero, J. Fierrez, J. Galbally, E. Maiorana, and P. Campisi, "Implementation of fixed length template protection based on homomorphic encryption with application to signature biometrics," in *Proc. Conf. on Computer Vision and Pattern Recognition Workshops (CVPR)*, 2016, pp. 191–198.
- [284] M. Gomez-Barrero, C. Rathgeb, J. Galbally, C. Busch, and J. Fierrez, "Unlinkable and irreversible biometric template protection based on Bloom filters," *Information Sciences*, vol. 370–371, pp. 18–32, 2016.
- [285] M. Gomez-Barrero, E. Maiorana, J. Galbally, P. Campisi, and J. Fierrez, "Multi-biometric template protection based on homomorphic encryption," *Pattern Recognition*, vol. 67, pp. 149–163, Jul. 2017.
- [286] M. Gomez-Barrero, J. Galbally, A. Morales, and J. Fierrez, "Privacy-preserving comparison of variable-length data with application to biometric template protection," *IEEE Access*, vol. 5, no. 1, pp. 8606–8619, Dec. 2017.
- [287] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *Proc. Advances in Cryptology — EUROCRYPT*, 1999, pp. 223–238.
- [288] H. Zhu, X. Meng, and G. Kollios, "Privacy preserving similarity evaluation of time series data," in *Proc. Intl. Conf. on Extending Database Technology (EDBT)*, 2014, pp. 499–510.
- [289] G. M. Penn, G. Pötzelsberger, M. Rohde, and A. Uhl, "Customisation of Paillier homomorphic encryption for efficient binary biometric feature vector matching," in *Proc. GI/IEEE Intl. Conf. of the Biometrics Special Interest Group (BIOSIG)*, 2014, pp. 121–132.

- [290] M. Barni, T. Bianchi, D. Catalano, M. Di Raimondo, R. D. Labati, *et al.*, "A privacy-compliant fingerprint recognition system based on homomorphic encryption and fingercode templates," in *Proc. IEEE Intl. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, IEEE, 2010, pp. 1–7.
- [291] X. Anguera and J. F. Bonastre, "A novel speaker binary key derived from anchor models," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2010, pp. 2118–2121.
- [292] J. F. Bonastre, P. M. Bousquet, D. Matrouf, and X. Anguera, "Discriminant binary data representation for speaker recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 5284–5287.
- [293] G. Hernández-Sierra, J. F. Bonastre, and J. Calvo de Lara, "Speaker recognition using a binary representation and specificities models," in *Proc. Iberoamerican Congress on Pattern Recognition (CIARP)*, 2012, pp. 732–739.
- [294] M. Paulini, C. Rathgeb, A. Nautsch, H. Reichau, H. Reininger, and C. Busch, "Multi-bit allocation: Preparing voice biometrics for template protection," in *Proc. Odyssey 2016: The Speaker and Language Recognition Workshop*, 2016, pp. 291–296.
- [295] Z. Erkin, M. Franz, J. Guajardo, S. Katzenbeisser, I. Lagendijk, and T. Toft, "Privacy-preserving face recognition," in *Proc. Intl. Symposium on Privacy Enhancing Technologies Symposium (PETS)*, 2009, pp. 235–253.
- [296] M. Barni, G. Droandi, and R. Lazzeretti, "Privacy protection in biometric-based recognition systems: A marriage between cryptography and signal processing," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 66–76, 2015.
- [297] J. Bringer, H. Chabanne, and A. Patey, "Privacy-preserving biometric identification using secure multiparty computation: An overview and recent trends," *IEEE Signal Processing Magazine*, vol. 30, no. 2, pp. 42–52, 2013.
- [298] A. K. Jain, L. Hong, S. Pankanti, and R. Bolle, "An identity-authentication system using fingerprints," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1365–1388, 1997.
- [299] T. Bianchi, S. Turchi, A. Piva, R. D. Labati, V. Piuri, and F. Scotti, "Implementing fingercode-based identity matching in the encrypted domain," in *Proc. IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS)*, 2010, pp. 15–21.
- [300] S. Ye, Y. Luo, J. Zhao, and S. Cheung, "Anonymous biometric access control," *EURASIP Journal on Information Security*, vol. 2009, p. 865 259, Nov. 2009.

- [301] Y. Luo, S. C. Sen-ching, and S. Ye, "Anonymous biometric access control based on homomorphic encryption," in *Proc. Intl. Conf. on Multimedia and Expo (ICME)*, 2009, pp. 1046–1049.
- [302] M. Blanton and P. Gasti, "Secure and efficient protocols for iris and fingerprint identification," in *Proc. European Symposium on Research in Computer Security (ESORICS)*. Springer Berlin Heidelberg, 2011, ch. Proc. European Symposium on Research in Computer Security (ESORICS), pp. 190–209.
- [303] D. J. Bernstein, J. Buchmann, and E. Dahmen, *Post-Quantum Cryptography*. Springer Science & Business Media, 2009.
- [304] M. Yasuda, T. Shimoyama, J. Kogure, K. Yokoyama, and T. Koshihara, "Packed homomorphic encryption based on ideal lattices and its application to biometrics," in *Proc. Intl. Conf. on Availability, Reliability, and Security (ARES)*, 2013, pp. 55–74.
- [305] —, "New packing method in somewhat homomorphic encryption and its applications," *Security and Communication Networks*, vol. 8, no. 13, pp. 2194–2213, 2015.
- [306] C. Patsakis, van J. Rest, M. Choraś, and M. Bouroche, "Privacy-preserving biometric authentication and matching via lattice-based encryption," in *Proc. Intl. Workshop on Data Privacy Management (DPM)*, 2015, pp. 169–182.
- [307] M. Pathak, S. Rane, W. Sun, and B. Raj, "Privacy preserving probabilistic inference with hidden Markov models," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 5868–5871.
- [308] M. Pathak and B. Raj, "Privacy preserving speaker verification using adapted GMMs," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2011, pp. 2405–2408.
- [309] —, "Privacy-preserving speaker verification and identification using Gaussian mixture models," *IEEE/ACM Trans. of Audio, Speech, and Language Processing (TASLP)*, vol. 21, no. 2, pp. 397–406, 2013.
- [310] J. Portêlo, B. Raj, A. Abad, and I. Trancoso, "Privacy-preserving speaker verification using garbled GMMs," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2014, pp. 2070–2074.
- [311] J. Portêlo, B. Raj, and I. Trancoso, "Logsum using garbled circuits," *Public Library of Science (PLoS One)*, vol. 10, no. 3, e0122236, 2015.
- [312] R. L. Rivest, L. Adleman, and M. L. Dertouzos, "On data banks and privacy homomorphisms," *ACM Foundations of secure computation*, vol. 4, no. 11, pp. 168–180, 1978.

- [313] C. Fontaine and F. Galand, "A survey of homomorphic encryption for nonspecialists," *EURASIP Journal on Information Security*, vol. 2007, p. 013 801, Dec. 2007.
- [314] J. Hoffstein, J. Pipher, and J. H. Silverman, *An Introduction to Mathematical Cryptography*. Springer, 2008.
- [315] T. W. Hungerford, *Algebra*. Springer Graduate Texts in Mathematics, 1974.
- [316] M. Bellare, A. Desai, D. Pointcheval, and P. Rogaway, "Relations among notions of security for public-key encryption schemes," in *Proc. Advances in Cryptology (CRYPTO)*, 1998, pp. 26–45.
- [317] P. Paillier and D. Pointcheval, "Efficient public-key cryptosystems provably secure against active adversaries," in *Proc. Advances in Cryptography — ASIACRYPT*, 1999, pp. 165–179.
- [318] IEEE Standards Association, *754-2008 IEEE standard for floating-point arithmetic*, 2008.
- [319] B. Thorne, *Python Paillier*, [Online] <https://github.com/n1analytics/python-paillier>, accessed: 2018-01-11, 2017.
- [320] E. Barker, L. Chen, A. Roginsky, and R. Davis, "Recommendation for pair-wise key establishment schemes using discrete logarithm cryptography," NIST, Tech. Rep. SP 800-56A Rev. 3, Apr. 2018, 56A.
- [321] M. I. Mandasari, M. Günther, R. Wallace, R. Saeidi, S. Marcel, and D. A. van Leeuwen, "Score calibration in face recognition," *IET Biometrics*, vol. 3, no. 4, pp. 246–256, 2014.
- [322] F. Kelly and J. H. L. Hansen, "Evaluation and calibration of Lombard effects in speaker verification," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 205–209.
- [323] —, "Score-aging calibration for speaker verification," *IEEE/ACM Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 12, pp. 2414–2424, 2016.
- [324] D. Castan, M. McLaren, L. Ferrer, A. Lawson, and A. Lozano-Diez, "Improving robustness of speaker recognition to new conditions using unlabeled data," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2017, pp. 3737–3741.
- [325] Z. Tan and M.-W. Mak, "I-vector DNN scoring and calibration for noise robust speaker verification," in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2017.
- [326] A. C. Yao, "How to generate and exchange secrets," in *Proc. IEEE Annual Symposium on Foundations of Computer Science*, 1986, pp. 162–167.

- [327] W. Henecka, S. Kögl, A.-R. Sadeghi, T. Schneider, and I. Wehrenberg, “TASTY: Tool for Automating Secure Two-party computations,” in *Proc. ACM Conf. on Computer and Communications Security (ACMCCS)*, 2010, pp. 451–462.
- [328] D. Demmler, T. Schneider, and M. Zohner, “ABY - a framework for efficient mixed-protocol secure two-party computation,” in *Proc. Network and Distributed System Security Symposium (NDSS)*, 2015.
- [329] Y. Ishai, J. Kilian, K. Nissim, and E. Petrank, “Extending oblivious transfers efficiently,” in *Proc. Annual Intl. Cryptology Conf. (CRYPTO)*, 2003, pp. 145–161.
- [330] G. Asharov, Y. Lindell, T. Schneider, and M. Zohner, “More efficient oblivious transfer and extensions for faster secure computation,” in *Proc. ACM SIGSAC Conf. on Computer & Communications Security (CCS)*, 2013, pp. 535–548.
- [331] Y. Polyakov, K. Rohloff, and G. W. Ryan, “PALISADE lattice cryptography library user manual,” Cybersecurity Research Center, New Jersey Institute of Technology (NJIT), Tech. Rep., Dec. 2017.
- [332] A. Sholokhov, M. Sahidullah, and T. Kinnunen, “Semi-supervised speech activity detection with an application to automatic speaker verification,” *Elsevier Computer Speech and Language (CSL)*, vol. 47, pp. 132–156, Jan. 2018.
- [333] A. Molina, A. Vergari, N. Di Mauro, S. Natarajan, F. Esposito, and K. Kersting, “Mixed sum-product networks: A deep architecture for hybrid domains,” in *Proc. Conf. on Artificial Intelligence (AAAI)*, 2018, pp. 3828–3835.
- [334] A. Vergari, A. Molina, R. Peharz, Z. Ghahramani, K. Kersting, and I. Valera, *Automatic Bayesian density analysis*, 2018. eprint: [arXiv:1807.09306](https://arxiv.org/abs/1807.09306).
- [335] H. Hao, “Deep learning in speaker recognition: I-vector classification using PLDA-RBM,” Master’s thesis, Danmarks Tekniske Universitet, 2015.
- [336] T. Stafylakis, P. Kenny, M. Senoussaoui, and P. Dumouchel, “PLDA using Gaussian restricted Boltzmann machines with application to speaker verification,” in *Proc. Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 2012, pp. 1692–1695.
- [337] D. H. Ackley and G. E. Hinton, “A learning algorithm for Boltzmann machines,” *Cognitive Science*, vol. 9, pp. 147–169, Jan. 1985.
- [338] P. Kenny, “Notes on Boltzmann machines,” Centre de recherche informatique de Montréal (CRIM), Tech. Rep., 2011.

- [339] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," in *Neural Networks: Tricks of the Trade*, Springer, 2012, pp. 599–619.
- [340] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.
- [341] T. Yamashita, M. Tanaka, E. Yoshida, Y. Yamauchi, and H. Fujiyoshi, "To be Bernoulli or to be Gaussian, for a restricted Boltzmann machine," in *Proc. IAPR IEEE International Conference on Pattern Recognition (ICPR)*, 2014, pp. 1520–1525.
- [342] D. E. Stansbury, *Matlab environment for deep architecture learning (MEDAL) vo.1*, [Online] <https://github.com/dustinstansbury/medal>, accessed 2015-09-22, 2013.
- [343] National Institute of Standards and Technology (NIST), "NIST 2018 speaker recognition evaluation plan," National Institute of Standards and Technology, Tech. Rep., 2018.
- [344] J. Katz and Y. Lindell, *Introduction to Modern Cryptography*, 2nd ed. Chapman & Hall/CRC Cryptography and Network Security, Nov. 2014.

EIDESSTATTLICHE ERKLÄRUNG

Hiermit erkläre ich, die vorgelegte Arbeit zur Erlangung des akademischen Grades (*Dr. rer. nat.*) mit dem Titel

Speaker Recognition in Unconstrained Environments

selbständig und ausschließlich unter Verwendung der angegebenen Hilfsmittel erstellt zu haben. Ich habe bisher noch keinen Promotionsversuch unternommen.

Antibes, den 28. Mai 2019

Andreas Nautsch

WISSENSCHAFTLICHER WERDEGANG

10/2008–03/2012	• Studium (B.Sc.) Computer Science Hochschule Darmstadt in Kooperation mit: atip GmbH, Frankfurt am Main
03/2012–04/2014	• Studium (M.Sc.) Computer Science Hochschule Darmstadt, in Kooperation mit: atip GmbH, Frankfurt am Main Masterarbeit: <i>Speaker Verification Using i-Vectors</i>
01/2013–06/2013	• Erasmus Mundus Programm, Mälardalens högskola, Västerås, Schweden
10/2013–08/2014	• Stipendiat, Deutschlandstipendium, CASED Stipendium, Hochschule Darmstadt
09/2014–12/2018	• Wissenschaftlicher Mitarbeiter, da/sec – Biometrics & Internet-Security, Hochschule Darmstadt, Nationales Forschungszentrum für angewandte Cybersicherheit
12/2015–03/2016	• Forschungsaufenthalt, ATVS Research Group, Universidad Autónoma de Madrid
03/2017	• Forschungsaufenthalt, Voxalys AB, Göteborg, Service Provider Forensic Phonetics
09/2017–12/2017	• Forschungsaufenthalt, Computer Vision Lab, University of Nottingham
seit 01/2019	• Wissenschaftlicher Mitarbeiter, Audio Security and Privacy Research Group, Digital Security Department, EURECOM

This document was typeset using the typographical look-and-feel *classicthesis* developed by André Miede. The style was inspired by Robert Bringhurst’s seminal book on typography “*The Elements of Typographic Style*”.

Hermann Zapf’s *Palatino* and *Euler* type faces (Type 1 PostScript fonts *URWPalladio L* and *FPL*) are used. The “typewriter” text is typeset in *Bera Mono*, originally developed by Bitstream, Inc. as “Bitstream Vera”.

The diagrams in this dissertation were created using *TIKZ*, *PGFPlots* by Christian Feuersänger, *Matlab*, *matlab2tikz* by Nico Schlömer, *matlabfig2pgf* by Paul Wagenaars, the *BOSARIS* toolkit by Niko Brümmer and Edward de Villiers, the *sidekit* by Anthony Larcher and Sylvain Meignier, *ECE plot* by Daniel Ramos, *bioplot* by Jos Bouten, and *validation toolbox* by Rudolf Haraksim—finally by Serge Rosmorduc’s *JSesh* 6.3.3, in honor of my adjourned field of study, Egyptology.



